

DOCUMENT RESUME

ED 143 674

95

TM 006 377

AUTHOR Irwin, Laura
TITLE Methodology for Aggregating Data in Educational Research. Final Report.
INSTITUTION American Univ., Washington, D.C.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
BUREAU NO NIE-4-1121
PUB DATE 76
NOTE 295p.; For related document, see ED 012 275
EDRS PRICE MF-\$0.83 HC-\$15.39 Plus Postage.
DESCRIPTORS Correlation; *Data Analysis; Models; Multiple Regression Analysis; *Research Methodology; Statistical Bias
IDENTIFIERS *Aggregation

ABSTRACT

This National Institute of Education report discusses problems of data aggregation and disaggregation which arise whenever the unit on which variables are measured is different from the unit of analytical, conceptual or policy-related interest. For example, researchers may want to know the effects of family background, of peer groups, or of school desegregation on individual students' achievement. Relevant data, however, may only be available for schools, even though the unit of analysis is the student. In situations such as this, the researcher must use findings based on the behavior of aggregates to make inferences about the behavior of individuals. This paper presents three methods for attempting to retain the advantages of aggregate data while minimizing the disadvantages. Detailed explanations of each of the three methods are provided along with a reanalysis of data from the Coleman study using some of the techniques presented. (Author/JKS)

* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

SP

TM006 377

FINAL REPORT

METHODOLOGY FOR AGGREGATING DATA
IN EDUCATIONAL RESEARCH

National Institute of Education Project No. 4-1121

Laura Irwin
Project DirectorThe American University
Washington, D.C.

1976

The research reported herein was performed pursuant to a grant contract with the National Institute of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to freely express their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official National Institute of Education position or policy.

TABLE OF CONTENTS

	Page
I. The Problems of Cost and Certainty	1
I.1 Introduction	1
I.2 Advantages of Microlevel Information	4
I.3 Problems in the Use of Microlevel Information	8
I.4 Summary	14
II. Effects of Data Aggregation: Bias and Uncertainty	16
II.1 Introduction	16
II.2 Aggregation Before Estimation	18
II.2.1 Grouping by the Values of the Independent Variable	23
II.2.2 Random Grouping	37
II.2.3 Grouping by Values of the Dependent Variable	41
II.2.4 Proximity Grouping	44
II.2.5 Misspecification of the Micromodel	49
II.2.6 The Expenditure Effects of State Aid: An Empirical Example of Alternative Grouping Strategies	51
II.3 Estimation Before Aggregation	65
II.3.1 Approach i: Perfect Aggregation	69
II.3.2 Approach ii: Aggregation Bias	74
II.3.3 General Critique of Traditional Econometric Approaches	93
II.3.4 An Application in Educational Research: The Expenditure Effects of Federal Aid	103

	Page
II.4	Summary and Conclusions 129
III.	Methods of Disaggregation: Effects on Bias 133
III.1	Introduction 133
III.2	General Approaches 135
III.2.1	The Method of Bounds 135
III.2.2	Ecological Regression and Correlation 140
III.2.3	Eliminating Bias in the Macrorelation: Multivariate Linear Ecological Regression 155
III.2.4	Eliminating Bias in the Macrorelation: Nonlinear Ecological Regression 160
III.2.4.1	Interaction: Variable Within-Unit Coefficients 166
III.2.4.2	Direct Effects: Invariant Within-Unit Coefficients 168
III.2.4.3	A Generalization: Types of Contextual Models 170
III.3	Ecological Regression and the Effects of Grouping 173
IV.	Grouping Strategies, Ecological Regression and Aggregation Bias 175
IV.1	Introduction 175
IV.2	Grouped Observations, Specification Error and Aggregation Bias--Correctly Specified Micromodels 178
IV.3	Grouped Observations, Specification Error and Aggregation Bias--Incorrectly Specified Micromodels 196
IV.4	Summary 215
IV.5	Implications 219

		Page
V.	Schools or Students: A Reanalysis of the Coleman Report	224
V.1	The Effects of Grouping: Standardized Measures, Contextual Effects, and the Specification of Models	226
V.2	Data Source and Measurement	233
V.3	Results	238
V.4	Summary and Implications	257
VI.	The Problem of Cross-Level Inference: Implications for Educational Research	261
	Bibliography	273

I. THE PROBLEMS OF COST AND CERTAINTY

I.1 Introduction

Problems of data aggregation and disaggregation arise whenever the unit on which variables are measured is different from the unit of analytical, conceptual or policy-related interest. Frequently, for example, in educational research the unit of analytical interest is the student himself. Researchers may want to know the effects of family background, of peer groups, or of school desegregation on students' academic achievement. Relevant data, however, may be available only for schools, even though the unit of analysis is the student. Similarly, researchers may want to know whether school districts respond to increments in state aid by increasing their own educational spending or by reducing (or failing to increase) their property taxes. However, the researcher may judge that the most reliable data on the fiscal behavior of school districts is available only in the form of state-wide averages.

In each of these situations, variables are measured on units larger than the unit of analytical interest. The researcher must use findings based on the behavior of aggregates to make inferences about the behavior of

individuals, be they students or school districts. Making these inferences is not a simple task. Data on aggregates are often more reliable, include more measures of relevant variables, and are much cheaper to obtain than data on individuals. However, aggregate data contain less information than individual level data; knowing only the mean income and achievement scores for a school reveals much less about students than knowing achievement scores and family income of each student (or even of a sample of students) in the school. Because of the information loss incurred by using aggregate data to make inferences about individuals, a researcher is less certain about the validity of his findings. Aggregate data have another advantage: they are easier to understand and assimilate. It is much easier to grasp the implications of a simple mean than a mass of data. Thus, the advantages of aggregate data--low cost, reliability, breadth, and clarity--are counteracted by the uncertainty of findings based on aggregate measurement when the unit of analysis is the individual.

Researchers often want, therefore, to retain the advantages of aggregate data while minimizing its main disadvantage--loss of information. There are three traditional ways of doing this, each of which constitute a topic of this study. First, altering how aggregate data are grouped can minimize the possibility of invalid inferences that may arise from using aggregate data to study individuals. While most existing aggregate data series are grouped by geographic

proximity, there are many other ways to group data. The advantages and disadvantages of alternative grouping strategies are discussed in section II.2 below.

Second, individual level data may be available, but need to be aggregated so that summary information can be produced. Altering the grouping process is a means of summarizing information before the information is used to derive estimates of individual behavior. However, a more propitious alternative to minimize information loss may be to derive individual level estimates and then summarize the individual level estimates themselves. There are, however, many ways these estimates can be summarized, and their advantages and disadvantages are discussed in section II.3 of the study.

Third, section III introduces and evaluates a variety of computational techniques which can be used to minimize the possibility of invalid inference normally associated with aggregate data.

The first and third of these traditional approaches--grouping strategies and computational techniques--are those most likely to be considered by educational researchers. The second approach is appropriate primarily for longitudinal data series uncommon in educational analysis. Consequently, section IV of this study integrates the first and third approaches by providing a general strategy which can be used in any given instance to determine whether using aggregate data to study individuals will produce valid or invalid conclusions. Finally, section V applies these results to a reanalysis of Coleman's study of student achievement.

Before considering each of these strategies in detail, it is well to mention briefly some other problems associated with the use of aggregate and individual level data. Aggregate information is ordinarily cheaper and easier to understand than detailed information about individuals, but detailed information ordinarily increases the researcher's certainty about the validity of his findings. However, there are additional considerations which should not be ignored in deciding whether to use aggregated or disaggregated information. These constitute the focus of the remainder of this introductory section.

1.2 Advantages of Microlevel Information

A variety of estimation and testing problems are associated with the use of highly aggregated data (Orcutt, 1968; Hanushek, Jackson and Kain, 1974). First, complete specification of relations is often easier at the micro than at the macro level, because microlevel theories in the social sciences are better developed. Perhaps this is because the social scientist ". . . is an individual, and he lives at a micro level" (Orcutt, 1968:75). Additionally, specification of causal effects of alternative policies is likely to be more attainable with microlevel data than with macro-level data. Microdata permit repetitive observations of responses to a set of stimuli in a wide variety of circumstances. For example, microdata in educational research can strengthen the validity of inferences about the effects

of alternative teaching techniques by producing observations for different kinds of students in different schools, with different teachers, and the like. Highly aggregated data produce, in effect, only one observation at any single point of time. Finally, most available aggregate data, whether analyzed synchronically or diachronically, are autocorrelated and multicollinear. They may be poor measures of what is intended, because they are either insensitive to short-run developments or contain so many unmeasurable feed-back loops that the difficulty of identifying short-run developments is compounded.

These problems become even more intractable in policy research for several reasons (Orcutt, 1968). First, policy variables are rarely varied independently, making it difficult to evaluate the impact of their separate effects. Thus, changes in the formula for distributing federal aid to education in Title I of the Elementary and Secondary Education Act are likely to be accompanied by increases in the total amount of aid available to most school districts. Secondly, instruments of policy are changed in response to changes in the behavior the policy is intended to affect, thereby introducing particularly pernicious forms of simultaneous equation difficulties. For instance, programs intended to give compensatory advantages to the most educationally disadvantaged are likely to be reinforced when it appears that there are more disadvantaged students. An analyst insensitive to problems of reciprocal causation

might thereby infer that compensatory policies produce more disadvantaged students when in fact the reverse is true. Third, policy variables account for only a minor part of the variation in the behavior they are intended to affect, as Coleman's (1966) findings demonstrate with remarkable clarity. Thus, major sources of variation may remain uncontrolled. The effects of uncontrolled variables are even more damaging to systematic analysis because policy changes are likely to be slow and incremental rather than abrupt and drastic. Hence, the effects are likely to be small and difficult to identify even under the best of experimental circumstances. Finally, governments, in avoiding differential treatment of comparable classes of individuals, preclude the possibility of experimental evaluation with effective control groups. For example, all children within the class of children defined as disadvantaged can receive federal aid from Title I of ESEA. It is consequently impossible to compare the educational progress of disadvantaged recipients with disadvantaged non-recipients.

While each of these problems make systematic policy analysis and evaluation particularly troublesome, they can, to some extent, be mitigated by the use of microdata. As emphasized in preceding paragraphs, microdata do provide less collinearity, more independent replications, and fewer feed-back loops. It is acknowledged that microdata are costly. However, because of the deficiencies of relying exclusively on macrodata, and because decision makers

demand summary information, the importance of relying on microdata and microtheory to derive aggregate relations becomes even more crucial.

An additional problem is associated with the use of most available aggregate series. The territorial unit for which measures are available may not be theoretically relevant to the behavior of interest (Valkonen, 1969). It was noted above that most territorial data measure the behavior of units of general government and not special units like school districts. Another problem is the questionable comparability of aggregates, especially when territorial or governmental units provide information essentially by self-enumeration (Scheuch, 1966:138-145). The difficulties of standardizing accounting information across a wide variety of school districts are well known. The extent to which aggregate data are a valid representation of the collectivity they summarize is also questionable (Scheuch, 1966:146-148). The greater the internal variation of the property in question, the less representative is the summary statistic. In the case of measures of central tendency, their representativeness becomes even more doubtful when distributions are highly skewed. These problems are likely to be particularly troublesome for summary indicators of educational resources in school districts, such as measures of family income, adult education, teacher quality and the like. Such measures are

likely to be both heterogeneous and skewed, particularly in large urban and suburban school districts.

While these problems are likely to be encountered whenever aggregate data are used, their effects can be mitigated by using microlevel data sampled within contextual units. In particular, the problems of theoretical relevance and representativeness of aggregate measures can be avoided because more valid measures can directly be derived from the disaggregated data.

I.3 Problems in the Use of Microlevel Information

One source of problems in gathering microlevel information stems from the fact that data about individual units are likely to come from a sample rather than from a complete enumeration of units. Most sampling procedures typically used in the social sciences implicitly assume that individual and not contextual traits influence individual behavior. In policy research, this is clearly an untenable assumption, because policies are administered by various political, and thus contextual, units in order to influence the behavior of individuals within those units. In policy research, more than in other kinds of research, it therefore becomes critically important to design samples which permit evaluation of how context, or policy, influences individual behavior. There are two methods by which such samples can be designed, but the choice of one method forecloses certain advantages of the other.

One method is to include the policy-relevant context as an explicit element in stratifying the sample. For instance, schools and school districts are units through which educational policies are formulated and administered. Coleman (1966:551) was able to analyze the impact of variables measured at the school level on the behavior of students within the corresponding school because schools were a known probability sample within a larger probability sample of counties. Unfortunately, however, the decision to use counties rather than school districts as primary sampling units precluded measurement of the impact of school district policies on individual schools and students. School districts, not counties, are usually the relevant administrative unit in educational policy. However, because school districts are special districts rather than units of general government, demographic data from the Census are not available for the former. Without demographic data--particularly about school enrollment by race--Coleman was unable to select a sample of schools within school districts which would allow him to conveniently analyze the effects of racial composition of schools on students. Thus, in order to properly stratify the sample of schools, Coleman was forced to rely on sampling within the less-policy relevant unit of general government for which Census data were available.

While this is not a general problem in policy research, it is clearly a general problem in educational research, because education policy is generally administered by special units of government. Although the subset of 1970 census tract data that were mapped onto a sample of school districts may help to ease this dilemma somewhat, the sample of school districts used in the mapping project is not likely to be appropriate for all educational research purposes. Thus, at least in educational research, it will be difficult to include the administrative unit which carries out educational policy as an explicit element in the stratification design because there is insufficient data about special districts.

The decision to include the policy-relevant context as an explicit element in stratification has another pitfall: it is costly. As Barton (1968:1) points out, there is an inherent contradiction between most descriptive surveys and surveys designed to analyze the influence of contextual stimuli on a population. Most descriptive surveys try to reduce sampling error. To reduce sampling error as much as possible, interviews are scattered as much as possible to reduce the effects of clustering. In other words, it is not very efficient to take a large sample of respondents from the same locale, because similar people tend to cluster together. The fact that interviews are almost always clustered to save travel costs and time is regarded as a regrettable compromise of the ideal, and in the analysis

the clusters are ignored. Analysis of the effects of policy contexts on individual behavior, however, requires systematic sampling and explicit analysis of clusters. This raises the costs of survey methodology, which are quite high already. Unless the increment in survey costs is acceptable, there is likely to be a cost in descriptive accuracy with regard to larger populations if the sample is clustered. In other words, samples which are both representative of between-unit variation and large enough to capture variation within administrative units may entail the sacrifice of national or regional representativeness of the population so sampled (Linz, 1969).

In summary, the decision to sample individuals within another sample of administrative or contextual units has two probable disadvantages. First, in educational research particularly, demographic characteristics of the administrative unit, namely schools or school districts, from the Census are likely to be unavailable or insufficient, making within-unit stratified sampling extraordinarily difficult. Secondly, unless one's research budget is improbably generous, sampling within administrative units entails a sacrifice in the accuracy with which larger populations can be described.

These disadvantages are offset by the primary advantage of within-unit samples, which is the validity with which contextual effects (or policy consequences) can be inferred.

This very advantage is the primary disadvantage of the second strategy for generating sample data about the behavior of individuals. The second strategy is to match or combine existing administrative and census data with standard stratified cluster samples of the relevant population. In other words, the context variables come from sources other than the survey and are then ascribed to each respondent who appears in the sample and who happens to reside in that context. The primary advantage of this second strategy is one of the disadvantages of the first: at a given cost, it entails no loss of descriptive accuracy. Simply stated, this is a cheaper strategy. It is also a strategy which may require no new, specially designed surveys, but which can make use of existing conventional surveys and existing sources of administrative and census data.

Nevertheless, it is important to understand the drawbacks of this strategy. This strategy of course can be used only when the administrative unit in which the respondent lives can be identified. Less obvious but equally important is that conventional surveys do not have enough interviews per sample point to make statistically justified statements about these primary sampling units as social or policy contexts (Harder and Pappi, 1969:54; Valkonen, 1969:34). It is also necessary that a sufficient number of locations, or administrative units, be reflected in the conventional sample to capture the full range of administrative variation. While the incentive to spread interviews

in a conventional survey over as large a geographical area as possible is likely also to generate an adequate sample of administrative units, this adequacy is not guaranteed by the strategy of combining survey and administrative data while it is guaranteed by the alternative strategy of within-unit sampling. Moreover, conventional samples which do successfully capture the full range of administrative unit variation are also unlikely to have large enough within-unit N's for statistical generalization.

Finally, the importance and utility of microlevel information depends on the amount of within-unit variation in the variables of interest. If within-unit variability is minimal, neither having microdata to measure that variability nor seeking estimates of within unit variation from aggregate information will add anything to knowledge (Linz, 1969). Thus, if rural school districts or small school districts are homogeneous, it would not be worthwhile to worry about problems of data disaggregation that might be encountered in analyzing such districts. The utility of seeking individual level data is also likely to be questionable when little is known about the subject. For example, if little or nothing were known about the factors which affect student achievement levels, it would be very difficult to know what information about individual students should be collected. Distinguishing between relevant and irrelevant information would be extraordinarily difficult. In fact, undoubtedly for this reason, early studies of

pupil achievement were based on aggregate data. Perhaps studies of the effects of school desegregation on pupil achievement and on inter-racial attitudes among students are at such a preliminary stage that, given considerations of time and cost, aggregate information is, at this point, more useful than individual data.

I.4 Summary

Numerous problems other than those stemming from the process of cross-level inference itself are associated with the use of micro and macrolevel information. Relying on microlevel information is costlier because detailed sample data are expensive and time-consuming to obtain. Those costs can be reduced somewhat by ignoring contextual effects, but the contextual effects of different administrative units are of crucial interest in policy research. Thus, reducing the monetary costs of microdata imperils its utility in policy research. Relying on aggregate series represents a considerable cost reduction. It also represents a reduction in production and analysis time. To a policy-maker, timelines increase utility. Yet the saving of money and time through the use of aggregate information sacrifices validity in a number of respects: problems of collinearity, feedback and measurement are less tractable for aggregated than for disaggregated data.

Consequently, as an alternative to exclusive reliance on either aggregated or disaggregated information, it is important to understand the statistical consequences of the process of cross-level inference. Under certain circumstances, using aggregate data to study individuals can retain many of the advantages of aggregate data with little or no loss of information. Similarly, under certain conditions using individual data to study groups entails none of the liabilities of aggregate data while incorporating the advantages of detailed information about individuals. Understanding these conditions is therefore important if the advantages of both kinds of information are to be simultaneously realized.

II. EFFECTS OF DATA AGGREGATION:

BIAS AND EFFICIENCY

II.1 Introduction

The purpose of this chapter is to evaluate alternative methods by which information about microlevel behavior can be summarized in a statistically unbiased and efficient fashion. Each of these methods assumes that either cross-sectional or temporal data are available to estimate the behavior of microunits. For example, cross-sectional financial and demographic data measured on school districts can be used to estimate the fiscal responsiveness of 50 individual states to changing demographic circumstances. Or similar temporal data may be available for each of more than 66,000 school districts in the U.S. so that individual equations for each district can be estimated. The implications of such a vast amount of microdata should be clear: to ignore these data is to lose information, but to assimilate and understand the policy implications of 50 individual state estimates, much less 66,000 school district estimates, is to ensure that decision-makers will ignore the findings. Therefore, these data or findings must be summarized in some fashion, particularly if they are to be used by policy makers.

Although a variety of methods for data aggregation have been proposed, they can be sorted into two basic approaches. One approach aggregates by grouping the individual level values on each variable. It then computes sums or averages for each group, and derives a single estimate of aggregate behavior from the sums or averages. The other approach uses the individual level values of each variable to estimate microcoefficients and then uses simple or weighted sums or averages of these coefficients to derive a single estimate of aggregate behavior. It is thus possible either to aggregate the microlevel variables before estimating macrolevel relations or to estimate microlevel coefficients before aggregating them.

More specifically, assume that it is possible to estimate a set of $i = 1, \dots, N$ individual level regression equations from either temporal or cross-section data. Additionally, note that the use of regression techniques is appropriate whenever dependent variables are conceptually interval, because there are regression-based analogies for dichotomous and polytomous independent variables (Fennessy, 1968). Letting each individual level dependent variable y_i , measured over t subunits, be a function of an individual level independent variable x_i also measured over t subunits, gives $i = 1, \dots, N$ equations, each of which can be written:

$$y_{it} = a_i + b_i x_{it} + e_{it}$$

The first approach, aggregation before estimation, requires that the individual level variables be summed or averaged over t and that the resulting sums or averages be used to estimate an aggregate coefficient. For example, using averages gives the macrorelation:

$$\bar{y}_i = a + b\bar{x}_i + e_i$$

The second approach sums or averages each microcoefficient a_i and b_i to derive macrocoefficients. Each approach will be further elaborated below.

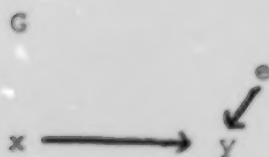
II.2 Aggregation Before Estimation

To aggregate the values of individual level variables before estimating aggregate relations first requires that each of the t subunits be allocated into i groups. How those groups are defined must be specified before the effects of aggregation on the bias and efficiency of aggregate estimates can be evaluated. There are four basic ways that the subunits t can be distributed among the larger units i in order to form groups for which the sums or averages are computed. First, the subunits t can be allocated randomly within each of $i = 1, \dots, N$ groups. Second, the subunits t can be allocated to each of $i = 1, \dots, N$ groups according to its value on the independent variable x_{it} . The microobservations with the highest values of x_{it} can be placed together in group $i = 1$; those with the next highest observations can be placed in group $i = 2$, and so

on until N groups (of as yet unspecified size) have been formed. Third, the subunits t can be allocated to each of $i = 1, \dots, N$ groups according to its value on the dependent variable y_{it} . And finally, each subunit can be allocated to one of N groups according to their temporal or spatial proximity. For example, if t represents time, all observations measured at time $t = 1$ are placed into the same group as are all observations measured at time $t = 2$, and so on for each unit of time. Alternatively, if t represents territorial subunits, all spatially contiguous subunits which lie within a larger unit are placed into the same group, and so on until each subunit has been placed into one of N larger units.

Each of these four grouping procedures can most readily be conceptualized using a causal perspective (Hannan, 1970:56-62; Blalock, 1964:95-114). In each case, the grouping criterion is a variable included in the model or related to a variable in the model, except in the case of random grouping. However, the error term is assumed to be unrelated to the grouping variable as well as to independent variables included in the model. (More complex cases will be considered in Chapter IV.) Simple inferences about the effects of alternative grouping procedures require that the relation between the grouping criterion and the variables in the model be made explicit. In the case of random grouping, individual level observations are assigned randomly to any one of N groups. In Figure 1, where G is the grouping criterion, there is no relation between G and the included variables. Since random assignment will have no disproportionate effects on the variance of

Figure 1: Random Grouping



one of the included variables relative to the other, it will have no effects on the value of the correlation coefficient (r_{xy}) or the regression coefficient (b_{yx}). Random grouping assures that these coefficients, whether estimated from micro or macroobservations, will remain unchanged.

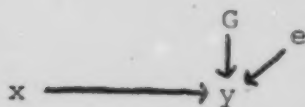
Grouping by the values of the independent variable x , however, increases the between-group variance in x relative to the within group variance, because all of the high values of x are put in one group, the next highest into another, and so on. Since x is assumed to be unrelated to other causes of y , represented by the error term e in Figure 2, the variation in x is increased relative to other sources of variation in y . Grouping by x disproportionately increases the variance of x relative to the variance of y . Consequently, R^2_{xy} , the proportion of variance in y explained by the variance in x , will be larger for grouped than for ungrouped data. In contrast, the value of the regression coefficient will remain unchanged because changes in the variance of the independent variable without concomitant changes in the variance of the dependent variable do not affect its value.

Figure 2: Grouping by Values of the Independent Variable



Grouping by values of the dependent variable affects both the coefficients of correlation and regression. In this case, the grouping criterion is systematically related to the dependent variable, as shown in Figure 3. Grouping will increase the between-group variance of y relative to its within group variance. However, this does not mean that the correlation computed from grouped data will be smaller than that computed from ungrouped data. The formula for the Pearson product moment correlation indicates rather that r_{xy} , and hence R^2_{xy} , behaves symmetrically with respect to x and y . R^2_{xy} will therefore increase whenever variation

Figure 3: Grouping by Values of the Dependent Variable

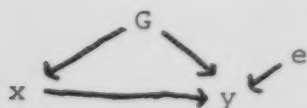


in either x or y is increased. In addition, the value of the regression coefficient will be larger when computed from data grouped by y than when computed from ungrouped data. This, however, is an artifact, and does not mean that using aggregates somehow changes the extent to which

a unit increase in x affects y . Since R^2_{xy} increases, and since b_{xy} must remain constant just as b_{yx} does when data are grouped by x , and, finally, since $R^2_{xy} = b_{yx}b_{xy}$, it follows that b_{yx} must increase over the value that would be obtained if the data were ungrouped.

The case of grouping units which are temporally or spatially proximate is intermediate between grouping on x and grouping on y , because it usually entails some of both, as shown in Figure 4. Units which have similar values on x or y are also likely to be spatially or temporally contiguous. The degree to which proximate units are more similar on one variable than another determines the degree to which the grouping is more closely related to the independent or the dependent variable. This in turn determines whether the values of R^2_{xy} and b_{yx} computed from grouped data in any given empirical instance will differ from the values computed from ungrouped data.

Figure 4: Proximity Grouping



The following four sections rigorously derive these results. Beginning with a simple bivariate model in which the error term is unrelated to the independent variable, random grouping and grouping according to the independent variable are shown to produce unbiased regression estimates. In contrast, grouping

by the dependent variable and by a variable related to both the independent and dependent variables produces biased estimates. An estimate is said to be biased if the mean of its sampling distribution is not equal to the value of the parameter being estimated. Thus, when individual level parameters are being estimated from aggregate data, the choice of an aggregation procedure determines whether estimates will be biased.

Similarly, the choice of an aggregation procedure determines whether correlations computed from aggregate data will equal their individual level counterparts. Only random grouping produces equal individual and aggregate correlations.

Finally, investigators are also concerned with the efficiency of regression estimates. Efficient estimates make it easier to distinguish between alternative models and increase the confidence with which predictions are made. The following sections show that, no matter how data are aggregated, grouped regression estimates are less efficient than their individual level counterparts. However, the efficiency loss that accompanies grouping is mitigated somewhat when data are grouped according to the independent variable.

II.2.1 Grouping by the Values of the Independent Variable

The most general method for logically deriving the effects of grouping on the bias and efficiency of regression and correlation coefficients is first to illustrate in detail the case of grouping by the independent variable, and then to show how other grouping strategies are special cases of this one. Following Cramer (1964), it will be demonstrated that grouping on x increases the value of R^2_{xy} , but does not bias the estimate b_{yx} , though it does

increase the variance and the standard error of that estimate by a small percent.

Consider a perfectly specified microregression:

$$y_{ij} = \alpha + Bx_{ij} + u_{ij} \quad (1)$$

where the x_{ij} are fixed, α and B are the unknown microparameters, and where $i = 1, \dots, t$ groups and $j = 1, \dots, n_i$ observations

within each group, so that $\sum_{i,j} n_{ij} = N$. The u_{ij} are independent random normal variates, with expected value of

zero and constant variance σ_u^2 . The u_{ij} are also independent of the x_{ij} . The mean of all observations

x_{ij} is denoted $\bar{x} = \frac{1}{N} \sum_{i,j} x_{ij}$ and the mean for each group of observations is $\bar{x}_i = \frac{1}{n_i} \sum_{j} x_{ij}$. The total sum of squared deviations from the mean of all observations is denoted by

$$s(x^2) = \sum_{ij} (x_{ij} - \bar{x})^2;$$

the between-group sum of squared deviations of each group mean from the mean of all observations is denoted by:

$$s(\bar{x}^2) = \sum_i n_i (\bar{x}_i - \bar{x})^2;$$

and the within-group sum of squared deviations of each observation from its group mean is denoted by:

$$s(x'^2) = \sum_{ij} (x_{ij} - \bar{x}_i)^2.$$

Three terms representing the covariation between the independent variable and the error term are defined below for the total, between-group and within group partitions respectively:

$$s(x, \mu) = \sum_{ij} (x_{ij} - \bar{x}) (\mu_{ij} - \bar{\mu})$$

$$s(\bar{x}, \bar{\mu}) = \sum_i n_i (\bar{x}_i - \bar{x}) (\bar{\mu}_i - \bar{\mu})$$

$$s(x', \mu') = \sum_{ij} (x_{ij} - \bar{x}_i) (\mu_{ij} - \bar{\mu}_i)$$

Using similar notation, the total covariation between each observation x_{ij} and y_{ij} is represented by

$$s(x, y) = \sum_{ij} (x_{ij} - \bar{x}) (y_{ij} - \bar{y}).$$

The total sum of squares or total covariation between x and y can also be partitioned into within group and between group components.

To show that grouping by the independent variable gives unbiased estimates of B , first substitute (1) into the cross-product $s(x, y)$, which gives:

$$s(x, y) = Bs(x^2) + s(x, \mu) \quad (2)$$

by the following steps:

$$\begin{aligned} s(x, y) &= \sum_{ij} (x_{ij} - \bar{x}) (y_{ij} - \bar{y}) \\ &= \sum_{ij} (x_{ij} - \bar{x}) [(\alpha + Bx_{ij} + \mu_{ij}) - (\alpha + B\bar{x} + \bar{\mu})] \\ &= \sum_{ij} (x_{ij} - \bar{x}) [B(x_{ij} - \bar{x}) + (\mu_{ij} - \bar{\mu})] \end{aligned}$$

$$\begin{aligned}
&= \sum_{ij} B(x_{ij} - \bar{x})^2 + \sum_{ij} (x_{ij} - \bar{x})(\mu_{ij} - \bar{\mu}) \\
&= Bs(x^2) + s(x, \mu)
\end{aligned}$$

Similarly, since $\bar{y}_i = \alpha + B\bar{x}_i + \bar{\mu}_i$ is the corresponding macroregression, it can be shown that:

$$s(\bar{x}, \bar{y}) = Bs(\bar{x}^2) + s(\bar{x}, \bar{\mu}) \quad (3)$$

The last terms in (2) and (3) are related by the partition:

$$s(x, \mu) = s(\bar{x}, \bar{\mu}) + s(x', \mu') \quad (3a)$$

However, because x and μ are independent, normally distributed and have zero means, it follows that the expected value of each term is zero:

$$\left. \begin{aligned}
&E(s(x, \mu)) \\
&E(s(\bar{x}, \bar{\mu})) * \\
&E(s(x', \mu'))
\end{aligned} \right\} = 0 \quad (4)$$

The variance of the first term in (4) can be expressed as follows:

$$\text{Var}(s(x, \mu)) = E\left[\sum_{ij} (x_{ij} - \bar{x})(\mu_{ij} - \bar{\mu}) - E\sum_{ij} (x_{ij} - \bar{x})(\mu_{ij} - \bar{\mu})\right]^2$$

Since the second term in brackets is zero, by equation (4), the expression reduces to:

$$\text{Var}(s(x, \mu)) = E\left[\sum_{ij} (x_{ij} - \bar{x})(\mu_{ij} - \bar{\mu})\right]^2$$

*This holds only in the case of grouping by x , as demonstrated below.

However, because x_{ij} and μ_{ij} are presumed to be independent:

$$\begin{aligned}\text{Var}(s(x, \mu)) &= E\left[\sum_{ij} (x_{ij} - \bar{x})^2 (\mu_{ij} - E(\mu_{ij}))^2\right] \\ &= s(x^2) \sigma_{\mu}^2\end{aligned}$$

Similar logic holds for each of the terms in (4), producing the following equalities:

$$\begin{aligned}\text{Var}(s(x, \mu)) &= s(x^2) \sigma_{\mu}^2 \\ \text{Var}(s(\bar{x}, \bar{\mu})) &= s(\bar{x}^2) \sigma_{\mu}^2 \\ \text{Var}(s(x', \mu')) &= s(x'^2) \sigma_{\mu}^2\end{aligned}\tag{5}$$

The final step needed to show the effects of grouping uses the fact that the two terms on the right hand side of the partition (3a) are independent. That is, the covariance of the between and within cross-products of x and μ equals zero:

$$\text{Cov}(s(\bar{x}, \bar{\mu}), s(x', \mu')) = 0\tag{6}$$

Consequently,

$$\begin{aligned}\text{Cov}(s(x, \mu), s(\bar{x}, \bar{\mu})) &= \text{Var}(s(\bar{x}, \bar{\mu})) \\ &= s(\bar{x}^2) \sigma_{\mu}^2\end{aligned}\tag{7}$$

by (5)*.

Each of these main results, indicated by numbered equations, will be used to show the effects of grouping by the independent variable, first on the bias of the regression coefficient computed from weighted group means and second on its efficiency. The first conclusion is that grouping by values of the independent variable does not bias the regression coefficient computed from group means. Since the regression coefficient computed from the microregression (1) is:

$$b = \frac{s(x, y)}{s(x^2)}$$

and that computed from the macroregression is:

* Because the within and between cross-products are independent, the only variation that $s(\bar{x}, \bar{\mu})$ shares with the total variation $s(x, \mu)$ is that contained in $s(\bar{x}, \bar{\mu})$. The formal proof of (7) follows:

$$\begin{aligned} \text{Cov}\{s(x, \mu), s(\bar{x}, \bar{\mu})\} &= \text{Cov}\{[s(\bar{x}, \bar{\mu}) + s(x', \mu')], s(\bar{x}, \bar{\mu})\} \\ &= E\{[s(\bar{x}, \bar{\mu}) + s(x', \mu')]s(\bar{x}, \bar{\mu})\} \\ &\quad - E[s(\bar{x}, \bar{\mu}) + s(x', \mu')]s(\bar{x}, \bar{\mu}) \end{aligned}$$

By (4), the term following the negative sign equals zero. Hence:

$$\text{Cov}\{s(x, \mu), s(\bar{x}, \bar{\mu})\} = E[s(\bar{x}, \bar{\mu})s(\bar{x}, \bar{\mu}) + s(x', \mu')s(\bar{x}, \bar{\mu})]$$

But by (6), the second term on the right hand side is zero. Therefore:

$$\begin{aligned} \text{Cov}\{s(x, \mu), s(\bar{x}, \bar{\mu})\} &= \text{Var } s(\bar{x}, \bar{\mu}) \\ &= s(\bar{x}^2)\sigma_{\mu}^2 \end{aligned}$$

$$\bar{b} = \frac{s(\bar{x}, \bar{y})}{s(\bar{x}^2)}$$

it follows, by substitution of (2) and (3) respectively into each of these numerators, that:

$$\begin{aligned} b &= \frac{s(x, y)}{s(x^2)} = \frac{Bs(x^2) + s(x, \mu)}{s(x^2)} \\ &= B + \frac{s(x, \mu)}{s(x^2)} \end{aligned} \quad (8a)$$

and

$$\begin{aligned} \bar{b} &= \frac{s(\bar{x}, \bar{y})}{s(\bar{x}^2)} = \frac{Bs(\bar{x}^2) + s(\bar{x}, \bar{\mu})}{s(\bar{x}^2)} \\ &= B + \frac{s(\bar{x}, \bar{\mu})}{s(\bar{x}^2)} \end{aligned} \quad (8b)$$

Taking expectations in both (8a) and (8b), and recalling from (4) that $E(s(x, \mu)) = E(s(\bar{x}, \bar{\mu})) = 0$, it is clear that:

$$E(b) = B \text{ and } E(\bar{b}) = B \quad (9)$$

Consequently, grouping by the values of x produces no bias in the estimate of the regression coefficient. Moreover, Feige and Watts (1972:345) show that this result also holds for regression equations with more than one independent variable.*

* This result, like the others in this chapter, presumes correct specification of the microrelation. It applies only to estimates of the independent influence of one independent variable on the dependent variable. In general, it does not apply to regression estimates of the effect of one independent variable on another.

In contrast, grouping by the independent variable does affect the variance of the estimate of the regression coefficient computed from the grouped data by increasing it. The reason is seen from a simple transformation of (8a) and (8b):

$$(b-B) = \frac{s(x, u)}{s(x^2)}$$

$$(\bar{b}-B) = \frac{s(\bar{x}, \bar{u})}{s(\bar{x}^2)}$$

so that, by (5):*

$$\text{Var}(b) = \frac{s(x^2)\sigma_u^2}{s(x^2)^2} = \frac{\sigma_u^2}{s(x^2)} \quad (9a)$$

$$\text{Var}(\bar{b}) = \frac{s(\bar{x}^2)\sigma_{\bar{u}}^2}{s(\bar{x}^2)^2} = \frac{\sigma_{\bar{u}}^2}{s(\bar{x}^2)} \quad (9b)$$

*The reasoning behind this derivation flows from squaring both sides and taking expectations of the transformed versions of (8a) and (8b). For (8a), the derivation is:

$$E[(b-B) - E(b-B)]^2 = E\left[\frac{s(x, u)}{s(x^2)} - E\left(\frac{s(x, u)}{s(x^2)}\right)\right]^2$$

On both sides of the equation, the term following the minus sign equals zero, since $E(b)=B$ and since $E(s(x, u)) = 0$. Therefore:

$$E(b-B)^2 = \frac{\text{Var } s(x, u)}{s(x^2)^2} = \frac{s(x^2)^{-2}}{s(x^2)^2}$$

by (5). Consequently,

$$\text{Var}(b) = \frac{s(x^2)\sigma_u^2}{s(x^2)^2} = \frac{\sigma_u^2}{s(x^2)}$$

The same logic holds for (8b).

Therefore, grouping by the independent variable will increase the variance of b by the ratio:

$$\frac{s(x'^2)}{s(\bar{x}^2)} \quad (10)$$

Conversely, grouping can be said to reduce the efficiency of the regression estimate by the ratio:

$$\frac{s(\bar{x}^2)}{s(x^2)} \quad (11)$$

(11) is always positive, because both terms are sums of squares, which are always non-negative. (11) is also always less than or equal to 1, because, since $s(x^2) = s(\bar{x}^2) + s(x'^2)$, $s(\bar{x}^2) \leq s(x^2)$. Thus, the closer the value of (11) is to 1, the smaller is $s(x'^2)$ and the less does grouping reduce the efficiency of the grouped estimate by increasing the variance of $\bar{b}$. Grouping by the independent variable does precisely this, because it maximizes between group variation, $s(\bar{x}^2)$, at the expense of within group variation on x (i.e., $s(x'^2)$). Grouping by the independent variable, then, does reduce the efficiency of the grouped estimate relative to the ungrouped estimate, but will not do so by very much.

In similar fashion, the estimate of the standard error of the regression coefficient computed from data grouped by values of x will exceed that computed from ungrouped data. At the individual level, the standard error is estimated by:

$$\hat{\sigma}_b = \sqrt{\frac{\sigma_v^2 / (N-2)}{s(x^2)}},$$

where v is the calculated residual and N is the number of individual observations. At the grouped level,

$$\hat{\sigma}_b = \sqrt{\frac{\sigma_{\bar{v}}^2 / (t-2)}{s(\bar{x}^2)}},$$

where t is the number of groups and $\bar{v}$ is the residual calculated from the aggregate level regression equation. The ratio of these estimates,

$$\frac{\hat{\sigma}_b}{\hat{\sigma}_b} = \sqrt{\frac{\sigma_v^2}{s(x^2)(N-2)}} \cdot \frac{s(\bar{x}^2)(t-2)}{\sigma_{\bar{v}}^2}$$

approximates:

$$\sqrt{\frac{(t-2)}{(N-2)}} \quad (12)$$

when data are grouped by the independent variable, for two reasons. First, grouping by the independent variable does not affect the variance of the residuals since x and μ are independent. Thus, the calculated residuals σ_v^2 and $\sigma_{\bar{v}}^2$ will be equal within sampling error. Second, grouping by the independent variable makes $s(\bar{x}^2)$ approach the value of $s(x^2)$. Expression (12) will always be less than unity, since t , the number of groups, is always less than N , the number of observations. Inspection of (12) also shows that the efficiency loss increases as the number of groups decreases.

Unlike the regression coefficient, the value of the correlation coefficient does not remain stable, but increases when computed from data grouped by values of the independent variable. However, instead of illustrating this effect by the ratio $\bar{R}^2/R^2$, an equivalent but algebraically more manipulable ratio, $\bar{C}/C$, will be used,

where $\bar{C} = \frac{\bar{R}^2}{1-\bar{R}^2}$ and $C = \frac{R^2}{1-R^2}$. To evaluate this ratio, consider the partition of the total variation in y :

$$s(y^2) = b^2 s(x^2) + s(v^2), \quad (13)$$

where $s(v^2)$ is the sum of squares of the computed residuals from the fitted regression line, not the true disturbances. This partition is a simple redefinition of R^2 , since dividing each term by $s(y^2)$ yields $1 = R^2 + \frac{s(v^2)}{s(y^2)}$, producing the familiar form of $R^2 = 1 - \frac{s(v^2)}{s(y^2)}$. Transformation of this expression gives another form of R^2 more useful for present purposes:

$$R^2 = \frac{b^2 s(x^2)}{b^2 s(x^2) + s(v^2)} \quad *$$

* The reasoning behind this expression follows from $R^2 = 1 - \frac{s(v^2)}{s(y^2)}$ by substitution.

$$R^2 = \frac{s(y^2)}{s(y^2)} - \frac{s(y^2) - b^2 s(x^2)}{s(y^2)}, \quad \text{by substitution of (13) into the numerator for } s(v^2)$$

$$= \frac{s(y^2)}{s(y^2)} - \frac{s(y^2)}{s(y^2)} + \frac{b^2 s(x^2)}{b^2 s(x^2) + s(v^2)}, \quad \text{by substitution of (13) into the denominator } s(y^2)$$

$$= \frac{b^2 s(x^2)}{b^2 s(x^2) + s(v^2)}$$

Since $1-R^2 = \frac{s(v^2)}{s(y^2)}$, it follows that

$$C = \frac{R^2}{1-R^2} = \frac{b^2 s(x^2)/b^2 s(x^2) + s(v^2)}{s(v^2)/s(y^2)}$$

Since the denominator in the numerator of this expression for C is equal to $s(y^2)$, the expression reduces to:

$$C = \frac{b^2 s(x^2)}{s(v^2)} \quad (14)$$

Similar formulae apply to weighted group means. Hence,

$$\frac{\bar{C}}{C} = \left(\frac{\bar{b}}{b}\right)^2 \frac{s(\bar{x}^2)}{s(x^2)} \frac{s(v^2)}{s(\bar{v}^2)} \quad (15)$$

Nothing about the value of $\bar{C}/C$ can be deduced in the general case. However, it has already been shown that when data are grouped by values of the independent variable,

$$E(b) = E(\bar{b}) = B \quad (16)$$

Consequently, since b and $\bar{b}$ will be within sampling error of the parameter B , the ratio $\bar{b}/b$ will be fairly close to 1. Moreover, since grouping by the independent variable is random with respect to the residuals v , the ratio $s(v^2)/s(\bar{v}^2)$ will also be close to one except for differences due to $N-2$ degrees of freedom for ungrouped data and $t-2$ degrees of freedom for grouped data. Thus the ratio reduces to approximately:

$$\frac{s(v^2)}{s(\bar{v}^2)} \approx \frac{N-2}{t-2} \quad (17)$$

Substituting (16) and (17) into (15) yields the ratio by which $\bar{C}$ will exceed C when data are grouped by the independent variable:

$$\frac{\bar{C}}{C} = \frac{s(\bar{x}^2)}{s(x^2)} \cdot \frac{(N-2)}{(t-2)} \quad (18)$$

(18) shows that there are two reasons why $\bar{C}$ is likely to exceed C . Looking at the second term on the right hand side of expression (18), it is obvious that as long as N , the number of observations, is larger than t , the number of groups, the term will be greater than unity. Moreover, as the number of groups decreases, each group will include a larger and larger number of observations. The ratio $(N-2)/(t-2)$ will also increase. Thus, even if the ratio $s(\bar{x}^2)/s(x^2)$ remains unchanged as the size of groups gets larger, the ratio $\bar{C}/C$ will still increase. That grouping data by the independent variable into groups of increasing size is alone sufficient to increase the aggregate estimate of R^2 has also been demonstrated empirically (Cramer, 1964:244-245; Slatin, 1969:900).

The first term on the right hand side of (18) also contributes to the increase of $\bar{C}$ over C when data are grouped by the independent variable. The reason for this was indicated in the previous section: grouping by the independent variable maximizes variation in the independent variable, but has no effect on the disturbances. Thus the statistical effect is to increase the grouped

correlation between x and y over the ungrouped correlation by increasing the ratio $s(\bar{x}^2)/s(x^2)$. Note however, since $s(x^2) = s(\bar{x}^2) + s(x'^2)$, this ratio will always be less than or equal to unity. Grouping makes $s(\bar{x}^2)$ approach $s(x^2)$, and, together with the second term, is enough to make $\bar{C}$ exceed C .

There are four basic conclusions to be drawn from this section. Each conclusion demands that the micro or individual level model is perfectly specified. However, if that demand is met, each conclusion holds for longitudinal and cross-sectional analyses, for dichotomous, polytomous or interval levels of measurement, and for bivariate or multivariate models. The first conclusion is that grouping by values of the independent variable yields unbiased estimates of the regression coefficient b_{yx} . Second, grouping by the independent variable increases the variance of the estimate of the regression coefficient by the ratio $s(x^2)/s(\bar{x}^2)$ and thus reduces the efficiency of that estimate by the inverse $s(\bar{x}^2)/s(x^2)$. This reduction, however, will be quite small, because grouping by x makes the value of this ratio approach unity, which represents no loss of efficiency. Third, grouping by x will increase the standard error of the regression estimate by the ratio $(N-2)/(t-2)$, resulting in a loss of efficiency equal to the inverse of that ratio. This loss becomes more severe as the size of groups increases or as the number of groups decreases. Fourth, as the standard error

increases by the ratio $(N-2)/(t-2)$, so will the value of the coefficient of determination increase by approximately the same amount. Thus, as grouping by values of the independent variable increases the precision by which each value of y is predicted from a given x , the precision by which the regression coefficient is estimated correspondingly decreases. Grouping by x , in fact, has been termed "optimal" because it produces unbiased regression estimates with little loss of efficiency. However, researchers with the option to group according to x should recognize that optimality occurs only when the aggregate model is properly specified.

II.2.2 Random Grouping

The effects of random grouping on estimates of regression and correlation are a simple extension of grouping by the independent variable. In the latter case, groups are formed by assigning individual cases to groups so as to maximize variation in the independent variable. In the former, individuals are assigned to groups randomly. In both cases, equation (4) holds. That is, the expected values of the total, within-group and between-group covariance of the independent variable and the disturbance terms are zero. However, random grouping will not affect the expected value of the between-group variation in x . Thus, by the partition $s(x^2) = s(\bar{x}^2) + s(x'^2)$, random grouping will make $s(x^2)$ approximately twice the value of $s(x'^2)$. These properties of random grouping have several consequences.

First, using exactly the same logic as in the case of grouping by the independent variable, it holds for random grouping that

$$E(b) = E(\bar{b}) = B.$$

Hence, random grouping does not bias estimates of the regression coefficient. Second, random grouping, unlike grouping by the independent variable, will not make the value $s(\bar{x}^2)$ approach the value $s(x^2)$. Consequently, equation (11), which measures the reduction in efficiency of the regression estimate, will not approach unity under random grouping. In fact, under random grouping, the ratio $s(\bar{x}^2)/s(x^2)$ will remain considerably less than unity, indicating that random grouping reduces the efficiency of the grouped estimate by increasing the variance of $\bar{b}$. This loss of efficiency will be greater than that due to grouping by the independent variable. Third, like data grouped by values of x , random grouping is also random with respect to the calculated residuals $s(\bar{v}^2)$ and $s(v^2)$. However, since the standard error of the regression coefficient is also affected by degrees of freedom, random grouping, like grouping by x , will increase the standard error by the ratio (12), which is $(t-2)/(N-2)$. Thus, under both kinds of grouping, as the size of each group increases, the number of groups decreases, and the standard error correspondingly increases.

The effect of random grouping on the value of the correlation coefficient computed from grouped relative to ungrouped data can be derived from equation (15):

$$\frac{\bar{C}}{C} = \left(\frac{\bar{b}}{b}\right)^2 \frac{\left(\frac{s(\bar{x}^2)}{s(x^2)}\right) \frac{s(v^2)}{s(\bar{v}^2)}}{1}.$$

It has already been shown that, under random grouping

$$E(b) = E(\bar{b}) = B$$

and

$$\frac{s(v^2)}{s(\bar{v}^2)} \sim \frac{(N-2)}{(t-2)},$$

and that the ratio $s(\bar{x}^2)/s(x^2)$ is likely to be a small positive number, considerably less than unity. Consequently, (15) reduces to

$$\frac{\bar{C}}{C} = \frac{s(\bar{x}^2)}{s(x^2)} \cdot \frac{(N-2)}{(t-2)}$$

The first term on the right of the equals sign will be a small number considerably less than one. The second term will be greater than unity, although how much greater depends on the number of groups, t , relative to the total number of observations.

Under the case of grouping by the independent variable, the ratio $s(\bar{x}^2)/s(x^2)$ is close to unity; hence $\bar{C}$ exceeds C to the extent that t , the number of groups, decreases.

However, under random grouping, the ratio $(N-2)/(t-2)$ is diminished by the ratio $s(\bar{x}^2)/s(x^2)$. Thus, $\bar{C}$ will exceed C much less under random grouping than under grouping by x . Random grouping will not have a large biasing effect on the correlation coefficient.

In conclusion, random grouping will not bias the estimate of the regression coefficient. It will have only a small biasing effect on the correlation coefficient, unlike grouping by the independent variable. However, it will reduce the efficiency of the regression estimate and increase its standard error even more than under data grouped by the independent variable. (For an empirical example, see Feige and Watts, 1972:354.) In effect, as far as efficiency of estimation is concerned, random grouping is "equivalent to taking a random sample of t observations from the N that are available and disregarding the rest; as a means of saving computational effort (this) procedure would be even better (than random grouping)" (Cramer, 1964:238). Thus, as a practical procedure for data aggregation, the gain afforded by random grouping over grouping by x in reducing the bias of the correlation coefficient is not worth the additional loss of efficiency and additional computational effort. The researcher would be better off had he simply taken a smaller sample in the first place.

II.2.3 Grouping by Values of the Dependent Variable

To group individual observations so that those observations with the highest values on the dependent variable in a perfectly specified model are placed in one group, those with the next highest in another, and so on, is to violate the independence between the disturbance term and the independent variable in the aggregate model. If it is assumed with no loss of generality that the perfectly specified micromodel is $y = a + Bx + \mu$, and that y and x and y and μ are positively related, then grouping the observations so that all high values of y are in one group will also result in observations which have high values of x and μ in the same group. As a result, in the aggregate model, high values of x will be associated with high values of the disturbance term. Such a model, of course, makes ordinary least squares an inappropriate procedure for obtaining unbiased estimates of the aggregate regression coefficient. The model produces biased estimates of the regression coefficient that would be obtained for ungrouped data. In other words, it will no longer be the case that $E(b) = E(\bar{b}) = B$. Additionally, because an aggregate model based on observations grouped by y violates the independence of $\bar{x}$ and $\bar{\mu}$, it will no longer be the case that $E(s(\bar{x}, \bar{\mu})) = 0$.

The fact that $E(b)$ no longer equals $E(\bar{b})$ means that, even if the micro model is correctly specified, permitting unbiased estimates in the ungrouped data, bias

will be introduced into the grouped estimator. This discrepancy between grouped and ungrouped estimates is in addition to the loss of efficiency in the grouped estimator due to reduction in degrees of freedom (Feige and Watts, 1972:346).

Finally, it is important to observe that grouping by y biases the grouped regression estimate in a positive direction if $\bar{x}$ and $\bar{\mu}$ are positively related. Examination of the second terms on the right hand side of (8a) and (8b) shows that the former disappears since $E(s(x, \mu)) = 0$. The latter remains positive since $E(s(\bar{x}, \bar{\mu})) > 0$. Hence it follows that $E(\bar{b}) > E(b)$. The opposite will be true if the covariation of $\bar{x}$ and $\bar{\mu}$ is negative.

The effect of grouping observations by the dependent variable on $\text{var}(b)$ relative to $\text{var}(\bar{b})$ and thus on the reduction in efficiency by which B is estimated, is difficult to determine. Assuming that the micromodel is perfectly specified, equation (9a) still holds:

$$\text{Var}(b) = \frac{\sigma^2}{s(x^2)}$$

However, because $E(\bar{b}) \neq E(b)$ and because $E(s(\bar{x}, \bar{\mu})) \neq 0$, equation 9(b) no longer holds. It thus becomes impossible in the general case to determine how grouping by the dependent variable will affect $\text{var}(b)$ relative to $\text{var}(\bar{b})$. While one might assert that this is an irrelevant problem because grouping by y biases the grouped estimator of B ,

it is also possible that there could be an efficiency gain together with only a small bias in the grouped estimator of B . Whether such conditions exist should be explored by statisticians, because resolution of the question could have considerable practical utility.

The standard error of the regression coefficient is a function of its variance and of the degrees of freedom. Grouping by y has indeterminate effects on the ratio of the variance of the grouped to the ungrouped estimator; but grouping also clearly reduces the degrees of freedom and thus reduces the efficiency of the grouped estimator at least by the ratio $(t-2)/(N-2)$. If grouping by y increases $\text{var}(\bar{b})$, there will be additional loss of efficiency. If, on the other hand, grouping decreases $\text{var}(\bar{b})$, the standard error of the grouped estimator could possibly be smaller than that of the ungrouped estimator.

The effect on R^2 , and thus on $\bar{C}/C$, of grouping which maximizes variation in the dependent variable is a mathematical artifact of the symmetry of the coefficient of determination. A grouping procedure which maximizes variation in either x or y will increase the correlation between x and y simply because $r_{xy} = r_{yx}$.

In conclusion, it is generally conceded that grouping procedures which maximize variation in the dependent variable of a perfectly specified micromodel are to be avoided because they are likely to reduce the efficiency of the grouped estimator by increasing its standard error.

Such grouping procedures also bias both the grouped regression and correlation coefficient in a known direction. However, the degree to which the variance of the grouped regression estimator and its standard error will be increased remains an open question. Hence, further consideration of the conditions under which there might be only a small loss of efficiency or even a gain in efficiency together with only a small biasing effect seems warranted.

II.2.4 Proximity Grouping

To group individual observations by proximity in a perfectly specified micromodel is to place observations into groups on the basis of either geographical or temporal proximity. For example, cross-section observations of a set of variables measured for individual school districts, counties or cities can be placed into separate groups based on whether units are located in the same state or not. Students can be placed into separate groups based on whether or not they live in the same school district or attend the same school. Similarly, a set of individual level observations measured over a period of time can be placed into groups based on whether each measurement was taken at the same time or not. Both kinds of grouping procedures will, to some extent, maximize variation in both the independent and the dependent variable. The reason for this is behavioral, not mathematical. Grouping data into years, states, counties and the like is not

likely to be random with respect to independent or dependent variables, because within each group, individuals are more likely to resemble each other with respect to either independent or dependent variables than they are likely to resemble people outside their group. For example, prices paid by different individuals for the same commodity at the same time are likely to be more similar to each other than those paid at different times. Incomes of different individuals received at the same time may also be more similar than incomes received at different times, although this is only a behavioral possibility not a mathematical necessity. Similarly, the education, incomes, voting behavior and consumption patterns of individuals who live in the same county or school district are likely to resemble each other more than they will resemble the same properties of individuals who live outside of the county or district.

These examples also make clear that the effects of proximity grouping, whether spatial or temporal, depend on how the particular grouping factor relates to the variables included in the micromodel. In other words, given a single micromodel defined for a set of individual level observations, to group those observations into counties or states, or to group them into 1-year or 5-year intervals, may change how the grouping factor is related to the independent and dependent variables, and even to

the error term. The most general statement that can be made is that proximity grouping is unlikely to be random. More likely is that proximity grouping represents some combination of grouping by x and grouping by y . Because no general conclusions about the effects of proximity grouping on the bias and efficiency of estimators can be derived, the most practical conclusion follows from the foregoing discussions of grouping by x and grouping by y . If individual level data are available, and if the micromodel is well specified, it is possible to estimate empirically how a given grouping factor--be it temporal or spatial--is related to the independent and dependent variables in the micromodel. The foregoing discussion suggests that, as long as the investigator is reasonably confident that the micromodel is well specified, the proximity criterion that should be used both maximizes variation in the independent variable and is independent of the residuals in the microregression. The proximity criterion chosen should also produce small enough groups so that efficiency is not reduced inordinately due to both fewer degrees of freedom and greater collinearity (Orcutt, Watts and Edwards, 1968: 782).

Meeting some of these proximity criteria should not ordinarily be too difficult with cross-section data, for two reasons. First, the independent variables in most social science and educational models are likely to be indicators of socioeconomic status. Both theoretical considerations and indirect empirical evidence suggest

that people of similar status live near each other spatially. However, as the spatial grouping grows very large, that homogeneity is likely to decrease. For example, grouping students by schools is likely to increase their within-group homogeneity and their between-group heterogeneity in social status (independent) variables more than will school district or state groupings. School district groupings will also have the same effect relative to state groupings. Second, social theory and indirect empirical evidence also suggest spatial groupings are more likely to maximize variation in socio-economic independent variables than in behavioral dependent variables. In other words, people who live in the same area are more likely to be homogeneous in their socio-economic status than in their observed behavior. Consequently, spatial groups which maximize variation in x , are independent of the disturbance terms, and are not too large are unlikely also to maximize significantly variation in y at the same time.

Proximity grouping may therefore approximate grouping by values of the independent variable. To the extent that this approximation is close, proximity grouping will share the optimal properties of that grouping strategy, as long as the individual level model is well specified.

Finally, proximity grouping has two other desirable features. First, data grouped spatially are more useful for policy analysis than any other grouping.

Policy makers are concerned with evaluating program performance within and among geographic units because they are elected from such units. They may also be concerned with performance among, say, different income groups, but this concern is likely to be secondary. Second, while grouping by values of the independent variable may be statistically optimal, it may be less practically optimal than proximity grouping. The former grouping process necessitates using aggregate data grouped by different variables each time the independent variables change. A single data series grouped by proximity is likely to be more useful for multiple purposes than data grouped by independent variables which change from one study to another.

Although the theory and indirect evidence supporting these proximity criteria for grouping data temporally is much less convincing than for grouping data spatially, there is considerable supporting evidence. Many empirical analyses of the effects of both spatial and temporal groupings, using widely varying topics of study, including consumer behavior, juvenile delinquency, illiteracy rates and black/white income differentials, indicate that relatively small spatial or temporal groupings in effect maximize variation in the independent variable relative to variation in the dependent variable. Consequently, they result in only a small loss of efficiency in grouped regression estimates, do not bias regression estimates, but

do result in higher correlation coefficients.*

II.2.5 Misspecification of the Micromodel

Each of the foregoing conclusions is based on the assumption that the underlying micromodel is perfectly specified, so that observations are independent of each other and so that no variable systematically related to both an included independent and dependent variable is omitted. Violation of this assumption makes it difficult to judge how alternative grouping procedures will affect the bias and efficiency of grouped estimators because misspecification entangles the effects of errors due to misspecification and errors due to data aggregation.

An example drawn from Blalock (1964:110) illustrates this difficulty with clarity. Suppose that z is an omitted variable which is also a cause of the included independent variable x and dependent variable y . If x and z are positively related then grouping together high values of x will also group together high values of z in a similar manner. Since z is also presumed to be a cause of y , the value of the ungrouped estimator of B_{yx} will reflect

* The studies upon which this conclusion is based are: W.S. Robinson (1950); Blalock (1964); Orcutt, Watts and Edwards (1968); Edwards and Orcutt (1969); Grunfeld and Griliches (1960); Feige and Watts (1972); Cramer (1964); and Slatin (1969). The only exception to this conclusion is found in the studies by Orcutt, some of which show a smaller aggregate R^2 . However, since how the temporal grouping procedure relates to the included variables is not specified, it is impossible to tell why this exception occurred.

the influence of z on y as well as that of x ; it will also reflect the influence of z on x . In addition to this specification bias in the ungrouped estimator will be bias due to aggregation. Because x and z are related, but z is omitted, grouping by x will no longer be random with respect to the omitted variables represented by μ .

Consequently, no longer will $E(s(x, \mu)) = E(s(\bar{x}, \bar{\mu})) = 0$.

Not only will $E(b) \neq B$, but $E(b) \neq E(\bar{b})$ (Feige and Watts 1972:346).

Additional verification of the consequences of misspecifying a micromodel is provided by Orcutt, Watts and Edwards (1968:784). They found that temporally aggregating a deliberately misspecified, simulated micromodel resulted in more biased and inefficient regression estimators than would be expected from aggregation effects alone.

The practical implications of misspecifying a micro-model are straightforward. Before aggregation, it is important to insure that all relevant variables are included in the micromodel. Although data aggregation always entails some information loss, that loss can be minimized by paying proper attention not just to the choice of aggregation procedures but also to the choice of the original micromodel. In addition, the last chapter of this study reveals the importance of properly specifying the macromodel. Sometimes, better specification of the macro than the micromodel can compensate for defects in

the microrelation. Thus, error due to aggregation does not always amplify errors due to specification, and can sometimes even reduce them.

II.2.6 The Expenditure Effects of State Aid: An Empirical Example of Alternative Grouping Strategies*

Understanding the effects of state educational aid on local school district spending is important. Whether state aid provides an incentive for school districts to increase their own educational spending or to replace local monies with state funds has crucial implications for the equity and efficiency of state aid formulas. Sacks (1972:153) finds, for example, that, controlling for other variables, suburban districts increase their local educational spending in response to state aid more than central city districts; neither government, however, uses state money to replace local revenues. If valid, this pattern suggests that state aid formulas only serve to increase existing financial inequities between central city and suburban schools. Moreover, to the extent that the benefits of public education are statewide rather than local, the costs should be borne by the state. It is possible that state funding formulas encourage some locales to overinvest in education, resulting in an inefficient allocation of scarce local revenues.

* This section was written with the assistance of Mark Shugoll, American University.

Commensurate with its importance, literature concerning the expenditure effects of state aid is extensive. Yet it is difficult to draw definitive conclusions because neither the statistical models nor the levels of aggregation are comparable. The purpose of this analysis is to demonstrate how differences in the level of aggregation affect conclusions concerning the expenditure effects of state aid. The analysis also demonstrates how the level of aggregation affects inferences about the impact of other variables on local educational spending. Finally, it shows that different grouping strategies result in inconsistent inferences.

To retain comparability of the variables, the model proposed by Sacks (1972) was selected for analysis because it contains indicators of most of the variables usually considered important in affecting the level of per pupil educational expenditures. Among the most significant of these variables is the income of persons residing in the school district. Income represents both demand and supply factors in educational spending. Higher income adults tend to demand higher levels of education for their children, in part because they themselves are likely to be more educated and in part because they can afford to spend more for education.

Income is only an imperfect indicator of supply at the school district level. It is a better indicator at

the state level. Local educational revenues are based on property taxes, and taxes depend on the assessed valuation of property and the tax rate. Income is only imperfectly correlated with assessed values and tax rates. Central city school districts, for example, tend to be "income poor" and "property rich" in contrast with outside central city districts which are "income rich" but "property poor." Using income instead of property taxes may therefore overrepresent the resources of outside central city areas and underrepresent the resources of cities. At the state level, income is a more adequate indicator of supply, because state revenues are based largely on income and sales taxes which are more perfectly associated with family income than property taxes. Unfortunately, the data source used in this study, the combined Census/ELSEGIS 1970 survey of school districts, contains no measure of district-wide property taxes. Moreover, since the purpose of the study is to compare the same models at different levels of aggregation, it is important to retain identical variables. Inclusion of district-wide property taxes would introduce an additional complication at the state level. In many states, state aid is nearly an exact inverse proportion of property values. Since state aid is included in the model due to its central theoretical role, it would be very nearly collinear with property values. Standard errors of the estimated regression coefficients would be very high. Even if the estimates were unbiased,

they would be unreliable and difficult to interpret (Christ, 1966:389). Therefore, one measure of supply used in this study is average family income. While median family income would be more appropriate, it could not easily be computed from the available data.

The enrollment ratio simultaneously encompasses several aspects of the demand for public education. This ratio is the percentage of the total population enrolled in public kindergarten, elementary and secondary schools. It reflects the demand for public education, but it also varies with the availability of nonpublic education and with the age distribution of the population. It is also sensitive to the proportion of the population attending secondary schools. The enrollment ratio will be higher where dropout rates are lowest, all other things being equal.

In addition to income, state aid per pupil measures another aspect of the supply of public education. While the actual impact of state aid on local school finances varies considerably from state to state and from district to district, its central theoretical importance for policy analysis has already been mentioned: if school monies are to be more equitably and efficiently spent, changing state aid formulas is likely to be the most feasible and direct means of accomplishing that goal. Changes in the distribution of federal aid to education would have only a

marginal effect, since federal aid is presently only about 7% of total educational expenditures, and dramatic increases in federal funding seem unlikely in the foreseeable future. Inclusion of federal aid per pupil in the models discussed below had no effect on the reported coefficients, and was consequently dropped from the analysis.

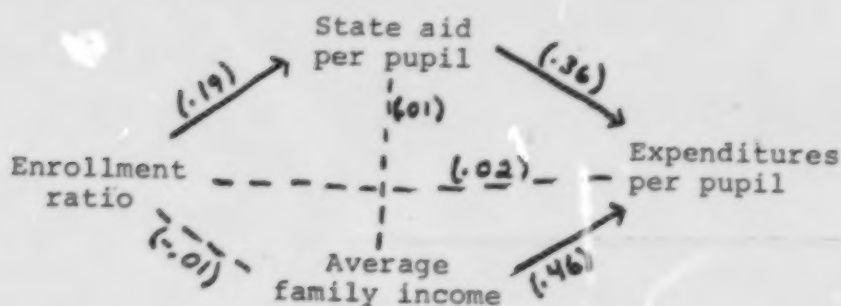
The results of the disaggregate analysis for the representative sample of U.S. school districts included in the Census/ELSEGIS data are shown in Table 1 and Figure 5. Figure 5 indicates the underlying causal or structural model which is most likely to have generated the coefficients reported in Table 1. At the disaggregate level, the enrollment ratio has virtually no direct effect on expenditures per pupil, as demonstrated in Figure 5 and

Table 1: Regression of Per Pupil Expenditures,
Disaggregate Model. (N=9707 school districts).

	Independent Variables		
	<u>State Aid per Pupil</u>	<u>Average Family Income</u>	<u>Enrollment Ratio</u>
Partial regression	.68	.05	-2.76
Standard error	.02	.00	.61
Beta	.36	.45	-.04
$R^2 = .34$			

Source: National Center for Educational Statistics,
Combined 1970 Census Fourth Count (Population)
School District Data Tapes and 1970 Elementary
and Secondary General Information Survey.

Figure 5: Causal Diagram of Per Pupil Expenditure Determinants, Disaggregate Model.



* Note: Figures in parentheses are simple correlations (r 's); dotted lines indicate where arrows are dropped.

confirmed by the partial regression coefficients in Table 1. The Beta coefficient is only $-.04$, by far the smallest of the reported coefficients. The regression coefficient indicates that a 1% increase in the number of children attending public schools brings about a decrease of about \$3.00 in per pupil spending, a seemingly negligible amount. Figure 5 indicates that the enrollment ratio does have a direct effect on state aid. This effect is attributable to the inclusion of district enrollment ratios in many state aid formulas.

Family income has a direct effect on school spending, lending empirical confirmation to the fiscal basis of recent state court decisions requiring equal educational spending across school districts*. Independent of income, state aid also has a large, direct effect on per pupil spending. However, that effect does not appear to be

*Of course, the effect of family income on school spending varies considerably from state to state. Individual state courts are therefore likely to reach different conclusions regarding the relation between income and expenditures.

stimulative, in contrast to Sacks' findings based on metropolitan areas. Among U.S. school districts, a \$1.00 increase in per pupil state aid increases total per pupil spending in a district by only \$.68. School districts probably use the remaining \$.32 to replace local monies they might otherwise have spent. Porter et al. (1973) and Wilensky (1970) suggest behavioral and theoretical explanations, respectively, for this observation.

Three alternative grouping strategies were applied to the disaggregate model, and, where appropriate, group sizes were varied. First, the data were grouped by geographic proximity; in short, the school districts were grouped by state, and the analysis was replicated using 51 states rather than the 9707 sample school districts as units of measurement. The resultant coefficients are reported in Table 2. In this analysis, grouping by proximity does not change the results significantly. With the exception of the regression coefficient for enrollment ratio, the coefficients are remarkably stable. Moreover, the standard errors do not increase, yet the proportion of explained variance does. If the estimates in the disaggregate model are unbiased, then the estimates in the state level model are likely to be unbiased also.

The instability in the estimate of the effect of enrollment on per pupil spending is difficult to explain, except by suggesting that the portion of the causal model

Table 2: Regression of Per Pupil Expenditures,
Aggregate Model, Data Grouped by State
(N=51).

	Independent Variables		
	<u>State aid per pupil</u>	<u>Average family income</u>	<u>Enrollment ratio</u>
Partial regression	.72	.07	9.83
Standard error	.01	.00	.50
Beta	.43	.65	.11
$R^2 = .71$			

Source: National Center for Educational Statistics,
Combined SDDT - ELSEGIS data, 1970.

relating the two variables is misspecified at either or both levels of analysis by the omission of a third variable. However, the coefficient relating state aid to local spending remains stable, and is the most relevant coefficient analytically. Thus, this analysis suggests that similar studies of expenditure effects could be carried out using state rather than school district data at considerably less cost and no sacrifice of certainty. In fact, it is possible that the state level model is even more valid than the district level model, because it may be better specified. Recall that income is probably a better indicator of supply at the state level, thereby improving the specification of the aggregate over the disaggregate equation. In some cases, then, it is possible

to attain greater certainty at less cost using aggregate data alone.

The school district data were also grouped by two of the independent variables included in the disaggregate model--state aid and income. The number of groups was also varied. The results of grouping the data into 44 state aid categories conform to expectation. (See Table 3.) With the exception of the enrollment ratio, whose coefficient is not significant, the coefficients differ little from those reported in Table 1, and the standard errors increase only slightly. Decreasing the number of groups, however, raises the standard errors enough so that both income and enrollment appear to have a statistically insignificant effect on per pupil spending. Predictably, the R^2 increases as the number of groups decreases. Also, the standardized Beta weights are much less stable than the unstandardized regression coefficients. For example, in Table 1, the standardized influence of income on spending is greater than that of state aid; the converse characterizes Table 3. Grouping on state aid maximizes its variation relative to other variables in the model, and thus makes the grouping variable appear more important when standardized measures are used. Overall, however, Table 3 suggests that using data categorized into 44 groups on the state aid independent variable loses little information. Roughly similar conclusions would be drawn from the

Table 3: Regression of Per Pupil Expenditures,
Aggregate Model, Data Grouped by State Aid

	Independent Variables		
	<u>State aid per pupil</u>	<u>Average family income</u>	<u>Enrollment ratio</u>
<u>N=44 **</u>			
Partial regression	.56	.05	4.37*
Standard error	.09	.02	11.70
Beta	.70	.24	.04
$R^2 = .53$			
<u>N=24 **</u>			
Partial regression	.52	.02*	19.62*
Standard error	.12	.03	20.18
Beta	.68	.11	.16
$R^2 = .59$			
<u>N=14 **</u>			
Partial regression	.43	.01*	38.48*
Standard error	.15	.04	27.15
Beta	.62	.09	.34
$R^2 = .67$			

* $P(b = 0) < .05$.

**See note, next page.

Source: National Center for Educational Statistics,
Combined SDDT - ELSEGIS data, 1970.

Notes for Table 3:

The N of 44 is based on 48 groups, each of which is defined in terms of \$25.00 increments. Of the 48 groups, 4 had no observations in them. The group with the lowest values includes all districts receiving between \$0 and \$24.99 in state aid per pupil. The group with the highest values includes all districts receiving between \$1375 and \$1399.99 per pupil. The number of observations in each of the 44 groups ranges from 1 to 1121.

The N of 24 is based on 25 groups, each of which is defined in terms of \$50.00 increments. Of the 25 groups, 1 had no observations in it. The number of observations in each of the 24 groups ranges from 1 to 2225.

The N of 14 is based on 14 groups, each of which is defined in terms of \$100.00 increments. The number of observations in each group ranges from 1 to 3750.

unstandardized coefficients in Table 1 and Table 3. However, in choosing between grouping by proximity and state aid, proximity grouping, at least in this instance, yields equally unbiased and slightly more efficient estimates. Proximity groupings are also likely to be more useful to administrators concerned with evaluation in their own geographic area of responsibility. Practicality, in this instance, would therefore dictate proximity grouping.

The results of grouping by another of the independent variables, family income, did not conform to expectation, as Table 4 illustrates. The influence of both income and

Table 4: Regression of Per Pupil Expenditures,
Aggregate Model, Data Grouped by Income

	Independent Variables		
	<u>State aid per pupil</u>	<u>Average family income</u>	<u>Enrollment ratio</u>
<u>N=32 **</u>			
Partial regression	.58*	.04	8.89*
Standard error	.63	.01	8.00
Beta	.12	.93	.13
$R^2 = .69$			
<u>N=14 **</u>			
Partial regression	-.23*	.03	-19.64*
Standard error	.74	.004	22.30
Beta	-.05	.87	-.11
$R^2 = .87$			

* $P(b = 0) < .05$

**See note on next page.

Source: National Center for Educational Statistics,
Combined SDDT - ELSEGIS data, 1970.

Notes for Table 4:

The N of 32 is based on 50 groups, each of which is defined by \$1000 increments ranging from \$0 to \$50,000. Of the 50 groups, 18 had no observations in them. The number of observations in each of the 32 groups ranges from 1 to 1912.

The N of 14 is based on 20 groups, each of which is defined by \$2500 increments ranging from \$0 to \$50,000. Of the 20 groups, 6 had no observations in them. The number of observations in each of the 14 groups ranges from 1 to 4168.

enrollment on spending are no longer significant. The only statistically significant variable is the grouping variable, income. Income also has the greatest impact on local spending according to the standardized Beta weights. In Table 3, where data were grouped by state aid, state aid had the largest standardized effect. Thus, grouping, by maximizing the relative variation of the grouping variable, makes that variable appear most important when only standardized measures are examined. The unstandardized effect of income, however, remains relatively stable. In the district level model, it is .05. In the aggregate model grouped according to income, it is .04 when 32 groups are used and .03 when 14 groups are analyzed. In contrast, the unstandardized effect of state aid does not remain stable. In the district level model, it is .68. When data are grouped by income, its coefficient is no longer significant. An unidentified source of misspecification in some portion of the district model undoubtedly accounts for the instability.

Finally, the district data were grouped by the dependent variable, expenditures per pupil. Grouping by the dependent variable leads to biased estimates of regression coefficients which are otherwise properly specified. Moreover,

the direction of bias can be determined: grouping by a dependent variable makes positive effects of an independent variable on the dependent variable appear more positive, and negative effects appear more negative. These patterns are demonstrated in Table 5 for state aid and for income. The effects of state aid on local school spending are nearly four times greater when aggregate data grouped by expenditure are used than when the disaggregate data are

Table 5: Regression of Per Pupil Expenditures,
Aggregate Model, Data Grouped by Expenditures

	Independent Variables		
	State aid per pupil	Average family income	Enrollment ratio
<u>N=45</u> **			
Partial regression	2.61	.15	-13.55*
Standard error	.47	.02	28.13
Beta	.39	.64	-.03
$R^2 = .85$			
<u>N=20</u> **			
Partial regression	2.48	.17	92.05*
Standard error	.71	.03	55.48
Beta	.37	.70	.13
$R^2 = .92$			

* $p(b = 0) < .05$

**See note, next page.

Source: National Center for Educational Statistics,
Combined SDDT - ELSEGIS data, 1970

Notes for Table 5:

The 45 groups are defined in terms of increments of \$30.00 except that the lowest group contains all school districts that spend between \$0 and \$419.99 per pupil and the highest group includes those spending between \$2269 and \$5560 per pupil. The number of observations in each group ranges from 36 to 670.

The 20 groups are defined by \$60.00 increments, except that the lowest group contains all districts spending from \$0 to \$479.99 per pupil and the highest group includes those spending between \$1916 and \$5560 per pupil. The number of observations in each group ranges from 176 to 1268.

used, as seen by comparing the regression coefficients in Table 1 with those in Table 5. Similarly, the effect of income is about three times greater in Table 5 than in Table 1. Once again, the enrollment ratio does not behave predictably, lending additional support to the likelihood that the portion of the model involving enrollment is poorly specified. Table 5 also displays the effects of decreasing the number of groups. Regardless of how data are grouped, standard errors as well as the proportion of explained variation increase.

Two important conclusions emerge from this analysis. First, in none of the aggregate models did the regression coefficient for the enrollment ratio behave as predicted. Similarly, when grouping by income, the state aid coefficient was not predictably behaved. Since all predictions about the effects of alternative grouping procedures assume that the disaggregate model is correct, failure of the predicted effects to materialize suggests that the original model is probably not valid. Thus, if a researcher actually has disaggregate data, using alternative grouping strategies to determine whether coefficients behave as they should is a means of testing alternative empirical models.

Second, due to the high costs of obtaining and ensuring the reliability and comparability of local school finance information, proximity grouping appears to be a reasonable and practical strategy for obtaining aggregate data, at least for investigations of the expenditure effects of state aid.

Grouping by the dependent variable yielded biased results, and grouping by the independent family income variable appeared unreasonable because it produced an unstable estimate of the effects of state aid. Grouping by state aid, another independent variable, yielded results which are probably unbiased but uses data which are clearly less practical administratively and less interesting to policy makers. Proximity grouping yielded similarly unbiased results. Additionally, statewide data are easier to obtain and are of considerable interest to state school officials. Since aggregate data are less costly to obtain than disaggregate data, statewide measures may suffice for the study of the effects of state aid on local spending. At least in this instance, proximity grouping at the state level loses little in validity at considerably less cost than disaggregated district finance data.

II.3 Estimation Before Aggregation

This section explicates the traditional econometric approach to problems of data aggregation. This approach, in contrast to a focus on grouping, is predicated on estimation of microrelations which are then aggregated in some fashion to form "consistent" macrorelations. Consistency has a very precise definition in this approach. It means that knowledge of the macrorelation and of the values of the independent macrovariables would lead to the same values of the dependent macrovariables as would

knowledge of the microrelations and of the values of the independent microvariables (Green, 1964:35). In other words, aggregation will be consistent "when the use of information more detailed than that contained in the aggregates would make no difference to the results of the analysis of the problem at hand" (Green, 1964:3). Both consistency and bias will be the criteria by which alternate aggregation schemes are evaluated.

More specifically, the consistency criterion can be illustrated by a simple example. There are two ways that a macrorelation can be derived. One is indirectly through a set of microrelations:

$$y_i = a_i + b_i x_i$$

The other is directly by the macrorelation:

$$y = a + bx$$

where $y = \sum_i y_i$ and $x = \sum_i x_i$ and $i = 1, \dots, N$ individual behavioral units. When the aggregate level variables are simple sums (or averages), substitution of the microrelations into the macrorelation will not yield consistent predictions of the dependent macrovariable.

To clarify, substituting the microrelation into the macrorelation gives:

$$\begin{aligned}
 y &= \sum_i y_i \\
 &= \sum_i (a_i + b_i x_i) \\
 &= \sum_i a_i + \sum_i b_i x_i
 \end{aligned}$$

But since the macrorelation has already been defined as:

$$\begin{aligned}
 y &= a + bx \\
 &= a + b \sum_i x_i
 \end{aligned}$$

it is evident that the two expressions for the dependent macrovariable, y , are not consistent, since

$$\sum_i b_i x_i \neq b \sum_i x_i .$$

Thus, in the traditional econometric approach, any aggregation scheme, such as that of simple summation used in this illustration, which leads to predictions from the microrelations which are inconsistent with those from the macrorelation is considered unacceptable. Moreover, virtually all of the literature using this approach centers on the derivation of those aggregation schemes which will yield information consistent with that inferred from microrelations alone.

The traditional econometric approach to aggregation problems can be further divided into two subapproaches.

The first assumes that a perfectly specified set of microrelations is given and that a macrorelation of the same form as each microrelation is posited. It seeks to find the relation between the macrovariables and the microvariables that must exist if the set of microrelations and the macrorelation are to be consistent. This approach is called "perfect aggregation." The second subapproach also assumes that a perfectly specified set of microrelations is given. It then posits an aggregation relation, and seeks the macrorelation which will be consistent with the given microrelation.

The remainder of this chapter will both explicate and critique each of these approaches in turn. It concludes with a critical evaluation of the consistency criterion, as well as with an empirical example of how the second approach can be used to resolve a dilemma of data aggregation common in studies of education finance. At the outset, however, it should be noted that both approaches have very unclear applications in cross-sectional studies. Both assume that time series data are available to estimate the behavior of individual units. Although some possible uses of these approaches in cross-sectional analyses will be suggested in the concluding sections, the explication which follows is restricted to situations where longitudinal microdata are available.

II.3.1 Approach i: Perfect Aggregation

This approach is easily set forth but its results are less readily interpretable. Assume that n perfectly specified individual microrelations are given:

$$y_{it} = a_i + b_i x_{it} + c_i z_{it} + u_{it} \quad (19)$$

where $i = 1, \dots, n$ individuals and $t = 1, \dots, T$ time periods. The variables y , x and z are identically defined in each of the n microrelations. Posit a macro-relation of the same form:

$$y_t = a + b x_t + c z_t + u_t \quad (20)$$

where, simply enough, the aggregate intercept, a , is assumed to be the sum of individual intercepts:

$$a = \sum_i a_i$$

and the macroslopes are assumed to be averages of the corresponding individual slopes:

$$b = \frac{1}{n} \sum_i b_i$$

$$c = \frac{1}{n} \sum_i c_i$$

Further assume that the aggregate dependent variable is a sum of the dependent microvariables:

$$y_t = \sum_i y_{it}$$

Given the first three assumptions, the macroequation (20) can be rewritten:

$$y_t = a + \left(\frac{1}{n} \sum_i b_i\right) x_t + \left(\frac{1}{n} \sum_i c_i\right) z_t + u_t \quad (21)$$

However, using just the fourth assumption, substitution of the microrelations (19) into (20) gives another macrorelation inconsistent with (21):

$$\begin{aligned} y_t &= \sum_i y_{it} \\ &= \sum_i a_i + \sum_i b_i x_{it} + \sum_i c_i z_{it} + \sum_i u_{it} \end{aligned} \quad (22)$$

These two reformulations of (20) are inconsistent because the second term in (21) is not equal to the second term in (22) and the third term in (21) is not equal to the third term in (22). What remains, then, is to define the aggregation relation between x_t and x_{it} and between z_t and z_{it} which will make these terms in (21) and (22) equal. Two conditions must be met for (21) and (22) to be equal:

$$\begin{aligned} (a) \quad \sum_i b_i x_{it} &= \left(\frac{1}{n} \sum_i b_i\right) x_t \quad (= \bar{b} x_t) \\ (b) \quad \sum_i c_i z_{it} &= \left(\frac{1}{n} \sum_i c_i\right) z_t \quad (= \bar{c} z_t) \end{aligned}$$

The only definitions of x_t in terms of x_{it} and of z_t in terms of z_{it} which render (19) and (20) consistent by

making (21) and (22) equivalent, are:

$$x_t = \frac{1}{\bar{b}} \sum_i b_i x_{it} \quad (23)$$

$$z_t = \frac{1}{\bar{c}} \sum_i c_i z_{it} \quad (24)$$

(See Allen, 1956:698.)* Using definitions (23) and (24) in the macrorelation (20) makes (19) and (20) consistent with each other. These definitions are called "perfect aggregation" (Allen, 1956:708).

While these definitions are a perfect mathematical solution to the problem of achieving consistent macro and microrelations, they are an imperfect substantive solution. Although the aggregate intercept and slopes are defined in an intuitively satisfying manner as the sum and average, respectively, of individual intercepts and slopes, the definitions of the independent macrovariables are less satisfying. To define an aggregate value as the sum of individual values weighted by the ratio of the individual to the aggregate slope has two consequences, each of which can best be explained in the context of an empirical example.

* The proof is simple, and is shown only for x .

$(\frac{1}{n} \sum_i b_i) x_t = \bar{b} x_t$ by definition. Substituting (23) gives:

$$\bar{b} x_t = \bar{b} \left[\frac{1}{\bar{b}} \sum_i b_i x_{it} \right] = \sum_i b_i x_{it} .$$

Suppose that the amount that any one of the 50 states spends per pupil in education over a period of years is a function of internal and external factors. The internal factor is the marginal propensity of state officials to spend each incremental dollar coming from the state's own revenue sources on education. The external factor is the marginal propensity of state officials to spend each extra federal dollar on education. (State officials can use federal funds earmarked for education as a substitute for state education monies, releasing the unspent local monies for non-education uses. Officials can also use non-educational federal aid as a substitute for local non-education monies, freeing these monies to be spent on education.) Thus, at the micro, or state level, per pupil state educational spending is a linear function of total state revenues per capita and of total federal revenues per capita.

To define an identical macroequation for the nation using (20), (23) and (24) has two effects. First, it gives greater weight to those states having the highest propensities to consume education. This results in a kind of double counting. Those states that, because of higher slopes, contribute more to the dependent microvariable, also contribute more to the value of the independent macrovariable. Second, it means that the value of both aggregate independent variables is "commodity specific"

(Allen, 1956:698). Specifically, each aggregate independent variable, even measured for the same period of time, takes on a different value if any other variables in the macrorelation are changed. For a given time sequence, the aggregate values of state revenues per capita and of per capita federal receipts by the states will depend on whether national spending patterns in education, highways, welfare, and the like are being examined.

These substantive consequences of perfect aggregation, may, in some contexts, be quite appropriate. For example, it is not necessarily undesirable to give those states with the highest marginal propensity to consume education greater weight in national estimates of state educational spending patterns, because those states have the greatest effect on national patterns. Unfortunately, however, the second consequence of perfect aggregation makes very little sense in this context. The amount of federal aid received by the 50 states in a given year does not change with each new dependent variable. Nonetheless, it is sometimes useful to let the values of independent macrovariables change with different dependent variables. Changing the dependent variable often implies that the independent variables have a slightly different meaning. This is not improbable if variables are considered to be imperfect indicators of more general but less directly measurable properties. For example, suppose interest

centers on family consumption of higher education as a function of family income. A commodity specific measure of aggregate income may be quite appropriate in this context, because the sociological meaning of income with respect to education is different than its meaning with respect to the consumption of dishwashers. While in the latter case, income may simply be an indicator of wealth alone, in the former it is an indicator of wealth as well as of occupational aspirations.

Thus, although no general rule can be formulated, there are some particular circumstances when perfect aggregation is a reasonable strategy for aggregating microrelations which are well-specified and temporally defined.

II.3.2 Approach II: Aggregation Bias

This approach also begins with properly specified microrelations (19). Unlike the first, it adopts simple aggregation:

$$Y_t = \sum_i Y_{it} \quad (25)$$

$$x_t = \sum_i x_{it} \quad (26)$$

$$z_t = \sum_i z_{it} \quad (27)$$

and derives a macrorelation of the form given by (20) whose coefficients yield predictions consistent with the

predictions that would be generated by the microequations. Thus the problem is to define the relation between the macroefficients a , b and c and the microcoefficients a_i , b_i and c_i which yield consistent predictions.

First, an auxiliary regression of each independent microvariable on the set of independent macrovariables is defined for each individual $i = 1, \dots, N$:

$$x_{it} = A_{xi} + B_{xi}x_t + C_{xi}z_t + v_{xit} \quad (28)$$

$$z_{it} = A_{zi} + B_{zi}x_t + C_{zi}z_t + v_{zit} \quad (29)$$

Then, substituting (28) and (29) into the microequation (19) and summing the results over all individuals, dropping the time subscripts for clarity, yields the desired macrorelation:

$$\begin{aligned} y &= \sum_i y_{it} \\ &= \sum_i a_i + \sum_i b_i A_{xi} + \sum_i c_i A_{zi} + \sum_i b_i B_{xi} x \\ &\quad + \sum_i c_i B_{zi} x + \sum_i b_i C_{xi} z + \sum_i c_i C_{zi} z \\ &\quad + \sum_i u_i + \sum_i b_i v_{xi} + \sum_i c_i v_{zi} \end{aligned} \quad (30)$$

which is of the form required by (20), where:

$$a = \sum_i a_i + \sum_i b_i A_{xi} + \sum_i c_i A_{zi} \quad (31)$$

$$b = \sum_i b_i B_{xi} + \sum_i c_i B_{zi} \quad (32)$$

$$c = \sum_i b_i C_{xi} + \sum_i c_i C_{zi} \quad (33)$$

$$u = \sum_i u_i + \sum_i b_i v_{xi} + \sum_i c_i v_{zi} \quad (34)$$

For greater clarity, (31), (32), (33) and (34) can be rewritten: *

$$a = \sum_i a_i + n \text{Cov}(b_i A_{xi}) + n \text{Cov}(c_i A_{zi}) \quad (35)$$

$$b = \bar{b} + n \text{Cov}(b_i B_{xi}) + n \text{Cov}(c_i B_{zi}) \quad (36)$$

$$c = \bar{c} + n \text{Cov}(b_i C_{xi}) + n \text{Cov}(c_i C_{zi}) \quad (37)$$

$$u = \sum_i u_i + n \text{Cov}(b_i v_{xi}) + n \text{Cov}(c_i v_{zi}) \quad (38)$$

* The proof of the equivalence of (31) with (35); of (32) with (36); of (33) with (37); and of (34) with (38) follows:

$$x_{it} = A_{xi} + B_{xi} x_t + C_{xi} z_t + v_{xit} \quad [\text{from (28)}]$$

$$x_{it} = x_t = \sum_i A_{xi} + \sum_i B_{xi} x_t + \sum_i C_{xi} z_t + \sum_i v_{xit}$$

Therefore:

$$\sum_i A_{xi} = \sum_i C_{xi} = \sum_i v_{xit} = 0$$

and $\sum_i B_{xi} = 1$ because the regression of a variable on itself is unity.

According to this approach, aggregation of micro-relations leads first to aggregation bias whenever the covariance terms in (35), (36), (37) and (38) do not disappear. If those terms do not disappear, aggregation is said to be biased because $E(b_i) \neq E(b)$ and $E(c_i) \neq E(c)$. Second, aggregation leads to inconsistent predictions of the dependent macrovariables, because substitution of (35), (36) and (37) into (20) would yield different values than would be predicted by the sum of the microrelations (19). Therefore, aggregation will be both inconsistent and biased whenever each independent microvariable is dependent on corresponding as well as on noncorresponding independent macrovariables.

$$z_{it} = A_{zi} + B_{zi}x_t + C_{zi}z_t + v_{zit} \quad [\text{from (29)}]$$

$$E_i z_{it} = z_t = E_i A_{zi} + E_i B_{zi}x_t + E_i C_{zi}z_t + E_i v_{zit}$$

Therefore:

$$E_i A_{zi} = E_i B_{zi} = E_i v_{zit} = 0$$

and $E_i C_{zi} = 1$ because the regression of a variable on itself is unity.

More specifically, there are three conditions under which aggregation will be both unbiased and consistent. First, if for each of the microrelations (19), $b_1 = b_2 = \dots = b_i$ for all $i = 1, \dots, N$ and $c_1 = c_2 = \dots = c_i$ for all $i = 1, \dots, N$, then the covariance terms in (35), (36), (37), and (38) will be zero, because the covariance of a variable with a constant is zero. In other words, aggregation will be unbiased and consistent whenever the influence of independent microvariables on the dependent microvariable is constant across all behavioral units.

$$a = \sum_i a_i + n\text{Cov}(b_i A_{xi}) + n\text{Cov}(c_i A_{zi}) \quad [\text{from (35)}]$$

$$= \sum_i a_i + \sum_i (b_i - \bar{b})(A_{xi}) + \sum_i (c_i - \bar{c})(A_{zi})$$

$$\text{since } \bar{A}_{xi} = \bar{A}_{zi} = 0$$

$$= \sum_i a_i + \sum_i b_i A_{xi} + \sum_i c_i A_{zi} \quad \text{since } \sum_i A_{xi} = \sum_i A_{zi} = 0.$$

$$b = \bar{b} + n\text{Cov}(b_i B_{xi}) + n\text{Cov}(c_i B_{zi}) \quad [\text{from (36)}]$$

$$= \bar{b} + \sum_i (b_i - \bar{b})(B_{xi} - \frac{1}{n}) + \sum_i (c_i - \bar{c})(B_{zi})$$

$$\text{since } \sum_i B_{xi} = 1 \text{ and } \sum_i B_{zi} = 0.$$

$$= \bar{b} + \sum_i b_i B_{xi} - \bar{b} \sum_i B_{xi} - \frac{1}{n} \sum_i b_i + n \frac{1}{n} \bar{b} + \sum_i c_i B_{zi} - \bar{c} \sum_i B_{zi}$$

$$= \sum_i b_i B_{xi} + \sum_i c_i B_{zi} \quad \text{since } \sum_i B_{xi} = 1 \text{ and } \sum_i B_{zi} = 0.$$

Second, if the coefficients in the auxiliary regressions (28) and (29) are constant and if those regressions are exact, then aggregation will be both consistent and unbiased. Both of those conditions are necessary. For bias to disappear, coefficients in the auxiliary regression must be constant so that the covariance terms in (35), (36) and (37) will equal zero. For consistency, the auxiliary regressions must be exact so that the covariance terms in (38) will equal zero.

Third, aggregation will also be unbiased and consistent if, even though the regression coefficients in (19) or in (28) and (29) are not constant over all individuals, all of the covariance terms in (35) through (38) are zero.

$$c = \bar{c} + n\text{Cov}(b_i C_{xi}) + n\text{Cov}(c_i C_{zi}) \quad [\text{from (37)}]$$

$$= \bar{c} + \sum_i (b_i - \bar{b}) (C_{xi}) + \sum_i (c_i - \bar{c}) (C_{zi} - \frac{1}{n})$$

$$\text{since } \sum_i C_{xi} = 0 \text{ and } \sum_i C_{zi} = 1.$$

$$= \bar{c} + \sum_i b_i C_{xi} - \bar{b} \sum_i C_{xi} + \sum_i c_i C_{zi} - \bar{c} \sum_i C_{zi} - \frac{1}{n} \sum_i c_i + n \cdot \frac{1}{n} \bar{c}$$

$$= \sum_i b_i C_{xi} + \sum_i c_i C_{zi} \text{ since } \sum_i C_{xi} = 0 \text{ and } \sum_i C_{zi} = 1.$$

$$u = \sum_i u_i + n\text{Cov}(b_i v_{xi}) + n\text{Cov}(c_i v_{zi}) \quad [\text{from (38)}]$$

$$= \sum_i u_i + \sum_i (b_i - \bar{b}) v_{xi} + \sum_i (c_i - \bar{c}) v_{zi}$$

$$\text{since } \sum_i v_{xi} = \sum_i v_{zi} = 0.$$

(Gupta, 1969, shows that, for this condition to be met, each covariance term must equal zero. It is insufficient simply for the sum of the two covariance terms in each equation to equal zero.)

It seems unlikely that these conditions will be met by empirical findings. The regression coefficients in (19) or in (28) and (29) are not likely to be constant across all behavioral units. Additionally, there are few theoretical reasons to expect the covariance terms to disappear. To illustrate from the example discussed earlier, if states with the highest marginal propensity to spend on education are also states whose own internal and external revenues increase most when total state and federal revenues are rising rapidly, then the covariance terms for (b_i, B_{xi}) and (c_i, C_{zi}) will be positive. Empirical evidence from the application in section II.3.4, in fact, corroborates this possibility.

$$= \sum_i u_i + \sum_i b_i v_{xi} - \bar{b} \bar{v}_{xi} + \sum_i c_i v_{zi} - \bar{c} \bar{v}_{zi}$$

$$= \sum_i u_i + \sum_i b_i v_{xi} + \sum_i c_i v_{zi} \text{ since } \bar{v}_{xi} = \bar{v}_{zi} = 0.$$

Similarly, suppose that Coleman's (1966) model of the dependence of pupil achievement on school facilities, family status and the social status of other students in the school has temporal validity. Among his findings was that school facilities had greater impact on disadvantaged than on advantaged students. However, disadvantaged students are also more likely to attend schools with poorer facilities, and these schools are likely to benefit least when school facilities are improving nationally. Since school facilities appear to have less impact on advantaged students who are most likely to attend schools with good facilities that are continually improved, the covariance term (b_i, B_{xi}) will be negative, not zero. Aggregation will be biased and inconsistent.

Green (1964) points out that aggregation can be biased but still consistent, in the sense that the error made in predicting the dependent macrovariable from the sum of the microrelations will be the same as that made in predicting from the independent macrovariables. This occurs when the auxiliary regressions are exact, such that:

$$x_{it} = A_{xi} + B_{xi} x_t + C_{xi} z_t$$

$$z_{it} = A_{zi} + B_{zi} x_t + C_{zi} z_t$$

Substitution of these expressions into (19), and summing the results will neither bring about the disappearance of aggregation bias nor lead to the same predictions from

(20) as from the summation of (19). But it will mean that the prediction errors made from (20) will be the same as would be made from summing (19):

$$\begin{aligned}
 y_{it} &= a_i + b_i x_{it} + c_i z_{it} + u_{it} \\
 &= a_i + b_i (A_{xi} + B_{xi} x_t + C_{xi} z_t) + c_i (A_{zi} + B_{zi} x_t + C_{zi} z_t) + u_{it} \\
 \sum_i y_{it} &= y_t \\
 &= \sum_i a_i + \sum_i b_i A_{xi} + \sum_i b_i B_{xi} x_t + \sum_i b_i C_{xi} z_t \\
 &\quad + \sum_i c_i A_{zi} + \sum_i c_i B_{zi} x_t + \sum_i c_i C_{zi} z_t + \sum_i u_{it} \\
 &= \left(\sum_i a_i + \sum_i b_i A_{xi} + \sum_i c_i A_{zi} \right) + \left(\sum_i b_i B_{xi} + \sum_i c_i B_{zi} \right) x_t \\
 &\quad + \left(\sum_i b_i C_{xi} + \sum_i c_i C_{zi} \right) z_t + \sum_i u_{it}
 \end{aligned}$$

which is of the form:

$$y_t = a + b x_t + c z_t + u_t$$

where

$$u_t = \sum_i u_{it}.$$

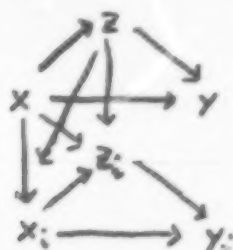
There is probably only one substantive application of this result. When the independent variables represent per capita (or per pupil) allocations of federal formula grants, the amount received by a given state is an exact proportion of the total amount available and will not be a

function of any other aggregate level variable. In these cases, it may be proper to write the auxiliary regressions as exact relations between independent micro and macrovariables.

The major problem with this second approach is the possibility of improper specification of the auxiliary regressions (28) and (29). Most discussions of these regressions suggest that they are either pure mathematical formalities with no substantive meaning or that their meaning is restricted to tracing the time path of microvariables relative to the movement of macrovariables (see for example Theil, 1954; Allen, 1956; Green, 1964. For an exception, however, see Peston, 1959-60). Nevertheless, these are inadequate interpretations, for if the auxiliary equations are not properly specified, estimates of their coefficients and of the source and amount of aggregation bias--defined as the covariance terms in (35), (36), and (37)--will be themselves biased. In short, the estimate of aggregation bias is dependent on proper specification not only of the microrelations, but also of the auxiliary relations.

The specification implicit in the foregoing equations can be diagrammed in the manner shown in Figure 6. The arrows linking x to z and x_i to z_i are simply intended to suggest that rarely does one encounter uncorrelated independent variables, particularly at the macrolevel

Figure 6: Macro, Micro, and Auxiliary Relations
(Model I).

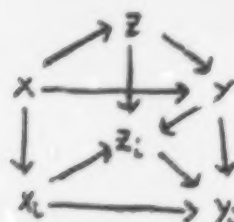


(Robinson, 1950; Duncan, Cuzzort and Duncan, 1961).

Although for many purposes this is an adequate representation of how aggregate variables may impinge on individual behavior, it is clearly not the most general representation. If the dependent macrovariable affects the relation between any of the microvariables, the expressions of the relation between macro and microefficients become more complex than those given in (35), (36), (37) and (38) above.

The empirical occurrence of a relation such as that illustrated in Figure 7 is not an improbability. In aggregating over commodities, for example, the existence of a budget constraint affects the relation between cost and demand for each commodity (Crecine, 1970). In aggregating over individuals, if the demand for certain kinds of goods or services is dependent in part on a "keeping up with the Joneses" phenomenon, then an individual's demand is in part a function of total demand for the item. In politics, this is called a bandwagon or underdog effect, and has been found to have differential effects on different

Figure 7: Macro, Micro and Auxiliary Relations (Model II).



people (Fleitas, 1971). Sociologists refer to the "structural effects" of contextual variables on individual behavior (Blau, 1960; Davis, Spaeth and Huson, 1961; Tannenbaum and Bachman, 1964). Many theories of deviant behavior, such as voting for an extremist candidate, or participating in legally or socially unacceptable acts, suggest that such behavior is in part a function of the availability of such activities to the individual. An individual who lives in an area in which voting for a left-wing candidate is traditional is more likely to vote for such a candidate than is his hypothetical twin who lives in an area where such voting patterns are rare (Soares and Hamklin, 1967). Similarly, a student attending a school where achievement levels are high may himself have higher achievement levels than his hypothetical twin attending a low achieving school. These examples suggest that aggregate dependent variables directly affect individual level variables and indirectly affect the relation between individual level variables.

Now consider these examples as they are represented in Figure 7. Note that the primary difference between Figure 7 and Figure 6 is that the aggregate dependent variable y affects both z_i and y_i in the former but not in the latter. In Figure 6, the aggregate independent variables are allowed to affect both corresponding and non-corresponding independent microvariables, whereas in Figure 7 the linkage with non-corresponding microvariables has been omitted for simplicity and without loss of generality. What then does Figure 7 imply about the relation between coefficients estimated from microdata and those estimated from macrodata?

First, to assume that microequation (19) is properly specified would be incorrect. If Figure 7 represents the theory, then equation (19) has an error term which includes an omitted variable, y , systematically related to the dependent variable y_i and to an independent variable, z_i , included in the equation. Consequently, it is necessary to rewrite equation (19), still assuming simple aggregation:

$$y_i = a_i + b_i x_i + c_i z_i + d_i y + e_i \quad (39)$$

It is also necessary to define three new auxiliary equations:

$$x_i = A_{xi} + B_{xi} x + v_{xi} \quad (40)$$

$$z_i = A_{zi} + C_{zi} z + D_{zi} y + v_{zi} \quad (41)$$

$$y_i = A_{yi} + D_{yi} y + v_{yi} \quad (42)$$

Note that (40) and (41) are similar to (28) and (29) respectively, except that in the former pair, the coefficients of non-corresponding aggregate independent variables are assumed equal to zero for simplicity and without loss of generality. In (41), the aggregate dependent variable is assumed to affect only z_i ; it could also affect x_i , but that would complicate arithmetic without changing the argument.

The three auxiliary equations can be substituted into the new microequations (39) which are then summed:

$$\begin{aligned} y &= \sum_i y_i \\ &= \sum_i A_{yi} + \sum_i D_{yi} y + \sum_i v_{yi} \\ &= \sum_i a_i + \sum_i b_i A_{xi} + \sum_i c_i A_{zi} + \sum_i b_i B_{xi} x \\ &\quad + \sum_i c_i C_{zi} z + (\sum_i c_i D_{zi} + \sum_i d_i) y \\ &\quad + \sum_i e_i + \sum_i b_i v_{xi} + \sum_i c_i v_{zi} \end{aligned} \quad (43)$$

The second line in (43) is no problem, for it simply expresses the regression of one variable on itself. Consequently,

$$\sum_i A_{yi} = \sum_i v_{yi} = 0$$

and

$$\sum_i D_i y_i = 1.$$

The third expression of (43), however, has y on both sides of the identity. Letting

$$(\sum_i D_i z_i + \sum_i d_i) = \lambda$$

and removing y from the right-hand side gives:

$$\begin{aligned} y = & [1/(1-\lambda)] (\sum_i a_i + \sum_i b_i A_{xi} + \sum_i c_i A_{zi}) \\ & + [1/(1-\lambda)] (\sum_i b_i B_{xi}) x \\ & + [1/(1-\lambda)] (\sum_i c_i C_{zi}) z \\ & + [1/(1-\lambda)] (\sum_i e_i + \sum_i b_i v_{xi} + \sum_i c_i v_{zi}) \end{aligned} \quad (44)$$

which is a macroequation of the proper form (20). However, these coefficients will be versions of those given in (31), (32), (33) and (34) weighted by $1/(1-\lambda)$. These coefficients will also be biased unless (39) rather than (19), and (41) rather than (29), are used to estimate them. Assuming that (39) and (41) are used, then if y_i tends to increase with y , so that $\sum_i d_i$ is positive; and if individuals for whom the effect of z_i on y_i is large

are also those for whom the aggregate dependent variable y has a positive effect on z_i , so that $D_{zi} > 0$ [since $\bar{D}_z = 0$, by (41)], then λ will be large. For $\lambda > 1$,

the weights will be negative and the estimates of aggregation bias given in (35) through (38) may not only have the wrong sign but also may be larger than the true estimates implied by (43).

Such a situation is not an impossibility. Theories of deviance suggest precisely this kind of phenomenon.

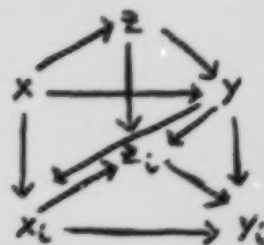
For example, if workers who have a high propensity to vote for left-wing candidates

are also most likely to live in or migrate to areas where left-wing voting is most common,

then $\sum_i \gamma_i D_{zi}$ would be large and positive. Since changes in the general level of left-wing voting are likely to affect positively only the voting habits of workers, while others would be affected negatively or not at all, then $\sum_i \gamma_i$ would be approximately zero. Hence one might estimate λ to be greater than unity. This is not, of course, to argue that one will always observe such a phenomenon, but only to suggest that the estimates of aggregation bias given by (35), (36), (37) and (38) are not perfectly general without also making the assumption that both the micro and auxiliary equations implied by Figure 5 are properly specified.

In Figure 7, the aggregate dependent variable affects only the relation between z_i and y_i . In Figure 8, a slight generalization of Figure 7 is offered, where y is allowed

Figure 8. Macro, Micro and Auxiliary Relations (Model III).



to affect directly each of the microvariables. With this addition, estimates of aggregation bias accordingly become more complex. Given the microequation (39), it is necessary to specify three new auxiliary regressions which follow from Figure 8:

$$x_i = A_{xi} + B_{xi} x + D_{xi} y + v_{xi} \quad (45)$$

$$z_i = A_{zi} + C_{zi} z + D_{zi} y + v_{zi} \quad (46)$$

$$y_i = A_{yi} + D_{yi} y + v_{yi} \quad (47)$$

Substituting these into (39) and summing the results gives a macroequation of the appropriate form:

$$\begin{aligned} y = & \sum_i \lambda_i + \sum_i \lambda_{bi} A_{xi} + \sum_i \lambda_{ci} A_{zi} \\ & + \sum_i \lambda_{bi} B_{xi} x + \sum_i \lambda_{ci} C_{zi} z \\ & + (\sum_i \lambda_{bi} D_{xi} + \sum_i \lambda_{ci} D_{zi} + \sum_i \lambda_{di}) y \\ & + \sum_i \lambda_{ei} + \sum_i \lambda_{bi} v_{xi} + \sum_i \lambda_{ci} v_{zi} \end{aligned} \quad (48)$$

Removing y from the right-hand-side of (48) means that

the estimates of aggregation bias given by (35), (36), (37) and (38) would be incorrect unless they were each weighted by

$$1/[1 - (\sum_i b_i D_{xi} + \sum_i c_i D_{zi} + \sum_i d_i)]$$

and had coefficients estimated by (39), (45) and (46).

This discussion is predicated on the belief that it is necessary to assume the auxiliary regressions are perfectly specified. To assume otherwise, and to additionally claim that the auxiliary coefficients can be estimated using ordinary least squares procedures is to accept biased estimates of aggregation bias itself, at least under certain circumstances. Properly specifying both auxiliary and individual relations yields estimates of aggregation bias different from those given in standard treatments of the problem, when an aggregate dependent variable affects the relation between variables measured at the individual level.

Similar biases in the estimates of aggregation bias itself given by (35) through (38) will arise if the micro-variables affect the macrovariables. The reverse is only assumed in each of the auxiliary regressions mentioned so far, yet the alternative assumption seems equally plausible in certain circumstances. For example, if the socio-economic status of students in a school affects the smoothness with which a school implements desegregation,

this is an implicit statement that individual properties affect a group property, and not the reverse. Similarly, it is less likely that the wealth of a state affects the relation between wealth and school spending in individual school districts, and more likely that the reverse is true. In other words, while the relation between district wealth and school spending may not differ from a wealthy state to a poor state, it is possible that the wealth of school districts within a state may have a stimulative impact on the state's propensity to spend on education.

These critiques are intended to indicate the conditions under which the approach to aggregation problems outlined in this section are appropriate. In sum, the approach outlined herein produces information about the amount and direction of aggregation bias and inconsistency, unlike any other approach mentioned. However, its applicability is limited to time-series models in which the micro and auxiliary equations are reasonably well specified, the dependent macrovariable has no direct effects on any of the microvariables, and the microvariables have no direct effects on any of the macrovariables. (For empirical applications, see Boot and deWitt, 1960; Gupta, 1969; and Irwin, 1975.)

II.3.3 General Critique of Traditional Econometric Approaches

Both of the approaches discussed in the preceding two sections of this chapter have one fundamental assumption in common--their preoccupation with the consistency criterion--and one apparent restriction--their applicability to temporal but not cross-sectional models. The purpose of this section is to discuss the implications of each of these problems.

The consistency criterion entails a definition of the aggregation problem as a comparison between predictions obtained from micro and macromodels (Peston, 1959-60). The criterion also implies the desirability of equivalence between aggregated predictions from the micromodel and those from the macromodel. Furthermore, the criterion usually assumes the logical primacy of microtheory. In other words, if the two predictions disagree, the aggregation procedure or the macrotheory, and not the microtheory, is viewed with suspicion. For example, Peston, in demonstrating that these approaches leave macrotheory somehow subsidiary to microtheory, cites Klein (1946), who states that the "behavior of individual households or firms. . . must form the basis of all theory of economic behavior." Arguing that macrotheory should be judged on its own terms and not in contrast to microtheory, Peston points out that it is invalid to infer that a theory of individual behavior must form the basis of all other theory. Moreover,

such a claim is an inaccurate portrayal of most economic theory anyway, because Klein is already talking of aggregates when he mentions households and firms.

What Peston saw as a problem fifteen years ago has more recently been characterized by Hannan (1970) as a conflict between the "consistency" or "homology" thesis versus the "discontinuity" thesis. The former takes macrotheory as a substitute for theories of individual behavior; consequently, any macroequation which is not a mirror of individual level equations is viewed as "biased." The latter thesis avers that different theories may account for the behavior of collectives and the behavior of individuals. Any disagreement between these theories may therefore reflect underlying social processes and not "bias" in the macrotheory.

The most contemporary view of this dilemma attests to the validity of the discontinuity thesis whenever individual behavior is a function not only of individual properties but also of contextual properties. The homology thesis is valid whenever individual behavior can be accounted for by individual properties alone. Thus, whenever both microvariables and macrovariables (or macrovariables alone) account for individual behavior, the discontinuity thesis is appropriate. Moreover, in this circumstance, it is inevitable that microequations and macroequations will be different. They will not be mirrors of each other.

For example, the equation

$$y_{it} = a_i + b_i x_{it} + e_{it}$$

might be a proper specification of the behavior of one individual unit measured over time. However, if the influence of x_{it} on y_{it} as estimated by b_i varies from one individual unit to another as a function of $\bar{x}_t$, then:

$$b_i = c_i + d_i \bar{x}_t .$$

The microequation which describes the behavior of all individual units taken together is therefore different:

$$\begin{aligned} y_{it} &= a_i + (c_i + d_i \bar{x}_t) x_{it} + e_{it} \\ &= a_i + c_i x_{it} + d_i x_{it} \bar{x}_t + e_{it} \end{aligned}$$

In addition, the corresponding macroequation is no longer linear:

$$\bar{y}_t = a + c\bar{x}_t + d\bar{x}_t^2 + e_t .$$

Thus, the presence of contextual effects means that individual and aggregate equations will no longer have exactly the same form.

This has relevance for educational research. Evidence from studies of pupil achievement strongly indicates that individual achievement is in part a function of the collective properties of other students with whom the individual attends school. Consequently, properly specified aggregate equations computed from school means will have a different form than equations representing the behavior of individual students.

The second problem with this approach is its apparent restriction to longitudinal models. Despite a number of statements to the contrary (e.g., Hannan, 1970; Grunfeld and Griliches, 1960), application of the technique used in the previous section to cross-sectional data is anything but straightforward. In fact, when data are aggregated over subunits of larger units, aggregation bias must necessarily disappear. Moreover, this result is identical to aggregating over time units rather than over individual units. In other words, summation over t rather than over i results, paradoxically, in the disappearance of aggregation bias, and aggregation over t in a longitudinal model is equivalent to aggregation over subunits in a cross-sectional model.

To illustrate, suppose there is a microrelation for each state in the U.S., whose coefficients are to be estimated on the basis of county data within each state

$$y_{ij} = a_i + b_i x_{ij} + e_{ij} \quad (49)$$

where $i = 1, \dots, n$ states and $j = 1, \dots, K_i$ counties within each state. Note additionally that the unit of analysis, county, and the grouping criterion, geographical proximity and political boundary, are immaterial to the results; individuals could also be the unit of analysis, grouped by income levels or by last names or by sex. The example is perfectly general. Suppose further that a

macroequation is sought:

$$y_i = a + bx_i + e_i \quad (i = 1, \dots, n) \quad (50)$$

where $x_i = \sum_j x_{ij}$ and $y_i = \sum_j y_{ij}$ and whose coefficients will be estimated from state data which are sums of corresponding county variables. Although most other studies of cross-sectional aggregation problems (e.g., Blalock, 1964; Robinson, 1950) use means rather than sums, the differences will not change the conclusions.

The macrovariables could have been formed by summing over i rather than j . This would be analagous to previous longitudinal results; it is also substantive nonsense, since it is equivalent to taking the first county in the first state, the first county in the second state, etc., and summing. This requires that the index number j have substantive meaning, as it does when it indexes a time sequence, and that each state have an identical number of counties. Consequently, the aggregation relation used here is one that could have been used had aggregation over time been at issue in the previous section. Equation (49) is a within-group regression and equation (50) is a between-group regression, to use terminology from previous analyses of aggregation problems in cross-sectional data.

An auxiliary regression of x_{ij} on x_i can also be defined for each state $i = 1, \dots, n$:

$$x_{ij} = A_i + B_i x_i + u_{ij} \quad (51)$$

This regression has different properties than the auxiliary regressions of the previous section. In equation (51), x_i is a constant vector, because for each county j in state i , the state sum is the same. (In previous models, for aggregation over individuals, and in this model, for aggregation over i , the values in the vector would vary.) The regression of a variable on a constant is zero, since their covariance is zero; hence $B_i = 0$ for all $i = 1, \dots, n$, and $A_i = \bar{x}_i = x_i$, while $\sigma_u^2 = S_{xij}^2$, and $\sum_j u_{ij} = 0$.

Summing over j what now remains of the auxiliary regression yields:

$$\begin{aligned}
 \sum_j x_{ij} &= x_i \\
 &= \sum_{j=1}^{K_i} A_i \\
 &= K_i A_i \\
 &= K_i \bar{x}_i \\
 &= K_i \left(\frac{1}{K_i} \right) \sum_j x_{ij} \\
 &= \sum_j x_{ij}
 \end{aligned} \tag{52}$$

In short, the auxiliary regression disappears, and there will be no aggregation bias in estimates of the coefficients in (50).

Even if a more realistic microequation is posited, the results do not change. Suppose, for instance, that

the effect of the independent microvariable on the dependent microvariable is influenced by an independent macrovariable:

$$y_{ij} = a_i + b_i x_{ij} + c_i x_i + e_{ij} \quad (53)$$

where $i = 1, \dots, n$ and $j = 1, \dots, k_i$. In addition, $x_i = \sum_j x_{ij}$ is a constant in each microequation but a variable in the macroequation, which will be misspecified if the influence of x_i on y_{ij} is omitted. Consequently, a macroequation of the form:

$$y_i = a + d x_i^2 + e_i \quad (54)$$

is required, where $d = (b + c)$, because the coefficients on the right of the equal sign in (53) cannot be separately identified.

The auxiliary regression is:

$$x_{ij} = A_i + D_i x_i^2 + u_{ij} \quad (55)$$

However, since the macrovariable is a constant, $D_i = 0$, and the whole equation will disappear for the reasons already cited.

In short, it appears that whenever this procedure is used to aggregate over subunits of a larger unit, there will always be constant independent variables in the auxiliary regression, which will make aggregation bias in the macroefficients disappear.

This conclusion is paradoxical, because there is sufficient empirical evidence to indicate that rarely will microcoefficients estimated from county data equal the coefficients computed from ecological or between-group regressions. Aggregation bias should not disappear, because it is unlikely that within-state regression coefficients computed from county (or school district) data will be equal for all states. In short, it is not a straightforward exercise to apply simply and formally the longitudinal results given in the previous section to cross-sectional data.

Suppose, however, that the cross-sectional model is subject to what Grunfeld and Griliches (1960) call a "synchronization" effect. This means that the value of one observation on an independent variable is influenced by the value of another observation on the same independent variable. Such an hypothesis is one way of formalizing Walker's (1969) notion of how innovations diffuse from one government to another; for example, educational innovations in one county (or school district, more properly) are more likely to be copied by some units than by others. However, it is necessary to hypothesize (though not to assume) which units are most likely to copy other units. In other words, time is a "natural" ordering, but individual behavioral units have no "natural" order. Thus, an ordering "rule" must be specified. One such "rule" is

that units are most likely to copy the behavior of the unit which is both geographically nearest and largest in population. This clearly suggests a positive synchronization effect. The microequation which reflects a positive synchronization effect is:

$$y_{ij} = a_i + b_i x_{ij} + c_i x_{i(j-m)} + u_{ij} \quad (56)$$

where $i = 1, \dots, n$ and $j = 1, \dots, k_i$. The second independent variable takes on the value of the county nearest and largest with respect to county j ; the former county has index number $(j-m)$. For some observations in some states, the value of $x_{i(j-m)}$ may be the same. This will be particularly true for those suburban counties which surround large central cities. Rural areas are likely to be subject to the value of $x_{i(j-m)}$ in neighboring suburban areas. There is no particular technical reason that the "cue-giver" could not be in a neighboring state; clearly on substantive grounds, that is reasonable.

To hypothesize the existence of synchronization effects in this manner does not seem far-fetched, in the light of Walker's (1969) evidence, and in the light of evidence that professional administrators are as likely to be influenced by other administrators in the same field in different units of government as they are by their own local constituents (Hofferbert, 1966). It is unlikely, in

fact, that states or counties or even school districts are truly independent units of analysis as implied by those who claim that local governments are autonomous political mini-systems (e.g., Dye, et al., 1963; Jacob, 1965).

If this belief is correct, then to omit a synchronous variable is to omit a variable which affects both the error term and an included independent variable; such a procedure would bias least-squares estimates in an unknown fashion.

Finally, it should be clear that the model as suggested in (56) is analogous to a longitudinal model with lagged independent variables. This analogy is crucial. Theil (1954) has shown that aggregating over time when at least one independent variable is lagged, which is equivalent to aggregating over individual units in cross-sectional models with synchronous effects, gives the same results for the source of aggregation bias in macroestimates as those in the previous section. Thus, the auxiliary regressions will not disappear. For example, forming the auxiliary regressions of the independent variables in (56) on corresponding and noncorresponding macrovariables gives:

$$x_{ij} = A_i + B_i x_i + C_i x_{i'} + u_{ij} \quad (57)$$

$$x_{i(j-m)} = D_i + E_i x_i + F_i x_{i'} + v_{ij} \quad (58)$$

$$\text{where } x_i = \sum_j x_{ij} \text{ and } x_{i'} = \sum_{(j-m)} x_{i(j-m)}$$

Although the first two terms on the right side of (57) and (58) will disappear when substituted into micro-equation (56) and then summed, the third term will not. (See Theil, 1954, pp. 40-43, for a demonstration.) Consequently, aggregation bias will not artificially disappear when the technique used in the previous section is applied to cross-sectional models with synchronization effects. In all other cross-section applications, the disappearance of aggregation bias is an artifact of the technique rather than a reflection of behavioral processes.

II.3.4 An Application in Educational Research: The Expenditure Effects of Federal Aid*

The purpose of this analysis is to show how aggregate information about the effects of federal aid on state spending can be broken into components. Analysis of the components permits assessing the sources of error in aggregated information without inordinately increasing the amount of raw information that must be assimilated. On the one hand, the empirical study shows how using individual information about each state is too massive and unwieldy for making inferences about the overall effects of federal aid. On the other hand, the study also shows how the use of aggregate data not only obscures much information about individual states but also leads to contradictory conclusions because there are many ways data can be aggregated. The method used to break aggregate information into components, however, provides a

* This section is an abridged version of Irwin (1975).

reasonable compromise between these extremes and seems novel as well as useful in the arena of policy choice.

Far from being academic, such a compromise can help assure local officials that federal officials have not ignored local variations. Many state and local officials charge that federal categorical grant-in-aid programs ignore individual variations to the extent of fostering unintended and undesirable consequences in many locales. For example, suburban spending is stimulated more by federal aid than central city spending (Campbell and Sacks 1967:151; Sacks, 1972). If aid is intended to equalize resources between suburban and central city governments, greater stimulation in suburban districts could hardly be considered a desirable or even intentional consequence, for it only reinforces existing disparities. In a similar vein, many local officials complain that uncertainty in the flow of federal aid necessitates spending it on hardware rather than on ongoing services. The validity of this complaint is likely to be greatest in communities least able to begin funding the service from local sources should federal revenues drop. A consequence of this uncertainty, then, is also to reinforce existing disparities.

More specifically, federal aid forms an important part of the context in which state officials must make spending decisions. On the average, more than 20 percent of state spending is funded by the federal government. While this amount varies from state to state and function to

function, it looms large in the budgetary calculus of state officials (Bahl and Saunders, 1965; Sacks and Harris, 1964). In fact, state officials both believe that federal funding should be increased and complain that federal funding increases their uncertainty in making revenue estimates. Thus, federal decisions about the amount of aid for a particular state function and about the formula by which it is distributed form an important part of the environment in which state officials operate.

Moreover, it seems reasonable to anticipate that grant consolidation will make federal and state politicians even more cognizant of the effects of federal aid on individual governments. Under consolidation, it will be less reasonable to assume that even if one federal aid formula is discriminatory, another will be favorable. State and local governments will find more at stake in a single formula. Federal officials who seek to satisfy diverse individual preferences but who ignore individual variations do so under greater peril with consolidation.

Consequently, federal officials will need to know more rather than less about the sources of local variations. At the same time, there seems little reason to believe that the costs of grasping the implications of large sets of numbers will decrease. In short, an increasing need for the certainty provided by disaggregated information will not be met by a decrease in the costs of assimilating and understanding such information. Thus, the stakes

involved in the trade-off between data aggregation and the risks of uncertainty are likely to increase. It therefore becomes important to understand more precisely the dimensions of this dilemma and to search for ways to mitigate it.

In order to depict the dilemma facing a public official who must choose between aggregated and disaggregated information, it will be necessary to assume rather than test the validity of an incremental theory of budgetary behavior. Much literature suggests that the high political and information costs of "rational" planning appear to necessitate strategies which minimize political conflict and comprehensive analysis. Incremental budgeting appears to be such a strategy. If in fact politicians generally try to minimize political conflict without suffering losses in their scope of authority, it follows that public budgets remain a fairly constant portion of private wealth. Expansions in the former come about automatically by expansions in the latter, rather than by increases in tax rates.

Similarly, agencies within a political jurisdiction seek to expand their own budgets without also expanding the scope of conflict. Hence, agency budgets are likely to remain a fairly constant proportion of a growing public budget. As the total public budget expands, so will the agency budget. Agency heads will often find this kind of expansion congenial, because it avoids conflict over

reallocation of scarce resources. Agency heads, of course, also look to the federal government for additional funds, regarding federal guidelines as a tolerable price of budgetary expansion.

Assuming that the incremental model is a reasonably accurate, even if simplified characterization of state agency budget behavior, it suggests that a regression equation constitutes an adequate formal analogue:

$$y_{it} = a_i + b_i x_{it} + c_i z_{it} + e_{it}$$

where $i = 1, \dots, 48$ contiguous states; $t = 1, \dots, 16$ years from 1952-1967; and

y_{it} = per capita expenditures for a specific function by the i th state in the t th year, excluding federal monies;

x_{it} = per capita federal aid for a corresponding specific function going to the i th state in the t th year; and

z_{it} = per capita income in the i th state in the t th year.

Application of this regression model to the three largest categories of federal aid--highways, welfare, and education--yields least squares estimates of the unknown parameters. The policy analyst or public official interested in how federal aid affects state spending decisions will pay particular attention to the estimated value of the second unknown, the partial regression coefficient "b". It shows how much states will increase per capita spending on

a given aided function if they receive one dollar per capita more in federal aid to that function, assuming that internal resources available to state officials remain constant.

The interpretation of "b" depends on the various matching provisions, if any, within each program. An estimated value greater than one in welfare, for example, which is generally characterized by a matching ratio of 1:1, means that a dollar per capita increment in federal aid makes the state spend more than required by the matching formula, even when the internal resources of the state are constant. The official might then conclude that federal aid to welfare stimulates state spending in welfare. If, on the other hand, the estimated value of the coefficient for welfare were less than one, the official might conclude that the state is spending less than required by the matching formula by using federal funds to replace what it otherwise might have spent from its own treasury. This is not necessarily budgetary chiseling by state welfare agencies, for federal matching requirements can be met by natural budgetary expansions from growth in the state resource base.

In highways, where matching ratios can be as high as 9:1, the federal official would be surprised to find estimated values of the partial regression coefficient which were very large. To meet such a matching ratio, a

one dollar per capita increment in federal aid requires little more than an eleven cent per capita state highway expenditure. A reasonable expectation is that such high matching ratios cause state highway agencies to use federal monies to replace what the state would have spent on highways in the absence of federal aid. Thus, federal highway aid may free state monies which would have gone to build roads for other uses. If so, a negative coefficient would be anticipated.

In contrast, federal aid to education is not generally characterized by matching provisions. A positive estimated value would suggest that federal aid stimulates state educational spending; a negative value would indicate that federal aid has replacement or substitution effects.

Assuming, then, the validity of the regression model, the data, and the interpretation of the estimated values, the federal official faces yet another set of dilemmas; should he derive the estimates from disaggregated or summary information? If he relies on the latter, how should he summarize? If the official decides to use disaggregated information, he could use the data for each of forty-eight contiguous states through 16 years to estimate a separate partial regression coefficient (b_1) in each of the three federal aid categories. This procedure yields $48 \times 3 = 144$ partial regression coefficients, each one of which communicates information about the effects

of a particular federal aid program on the spending decisions of a particular state, when the state's internal resources are constant. He would have to assimilate 144 pieces of information, yet he seeks only three conclusions. In using information about the effects of federal aid to consider new policy alternatives, even if he could assimilate 144 pieces of information, he must restrict his attention to uniform policies. He cannot have one policy for Alabama and another for California. In seeking summary information to make uniform policies and in order to avoid high information costs, he will therefore prefer three to 144 pieces of information: one coefficient for each federal aid category.

In choosing aggregate or summary information to develop uniform policy alternatives, it is necessary to decide how to summarize the raw data. A large number of options will appear reasonable: the data can be summed, it can be averaged, and any number of weighting schemes can be used in computing the sums or averages. At first only two simple options will be considered; a third will be introduced subsequently. One way to produce an aggregate picture is to pool the disaggregated information for all 48 states and all 16 years. The pooled data are then used to estimate a single regression for each federal aid category:

$$y_{it} = a' + b'x_{it} + c'z_{it} + e_{it} \quad (59)$$

where

$i = 1, \dots, 48$ contiguous states; and

$t = 1, \dots, 16$ years (1952-1967).

The values to be estimated, a' , b' , and c' , will be based on $16 \times 48 = 768$ observations. The federal official will be particularly interested in the estimated values of b' for each of the three federal aid categories.

Another simple way to aggregate is to use national aggregates for each of 16 years in a single regression for each aid category, where each observation is the sum of the data for the 48 states. Thus each variable represents a national sum for each of 16 years:

$$y_t = a + bx_t + cz_t + c_t \quad (60)$$

where:

$t = 1, \dots, 16$ years;

$y_t = \sum_i y_{it}$;

$x_t = \sum_i x_{it}$; and

$z_t = \sum_i z_{it}$.

In this case, the values to be estimated will be based on 16 observations.

The dilemma which now faces the policy analyst or public official is clearly illustrated in Table 6, which compares the estimated values for the partial regression

coefficients generated by the first regression equation using disaggregated information with the corresponding values estimated by the second using aggregated information. There are two aspects to the dilemma. The first is that using disaggregated and aggregated information to study the same problem can lead to different conclusions.

For example, the aggregate model for education shows that an extra dollar of federal aid effects only slightly more than a dollar increase in state spending. Yet the disaggregated information suggests that states increase their own educational spending by nearly four dollars for every federal dollar received. In the area of highways, where some matching ratios are as high as 9:1, some analysts have suggested that states would substitute federal dollars for state dollars. This hypothesis would be confirmed by the aggregate model, but contravened by the other. A federal welfare dollar appears to stimulate over two times more state dollars in the aggregate than in the disaggregate model. Thus, alternative hypotheses give contradictory results. Break (1967), for example, suggests that if federal aid were optimally distributed with respect to national interests, states should receive the greatest federal return on their dollar in areas having the greatest external benefits to other states. He suggests that welfare and education externalities are larger than highway externalities. Examination of the aggregate coefficients might lead one to conclude that,

on the basis of the three functions considered, which represent a large portion of federal aid, the distribution was reasonably optimal. Examination of the disaggregated data analysis would lead to a more ambiguous conclusion about the optimality of the existing distribution. Differences between the two models also appear with respect to the regression coefficient for income. Further, the differences between the models are not consistent: coefficients are not necessarily larger in one model than in the other.

Because one set of results is not clearly more "correct" than the other, it might be reasonable to choose those results in which one has the greatest confidence by selecting the model which minimizes the probability of a Type I error. The second part of the dilemma, however, is that even this strategy of minimizing error does not yield an unambiguous choice, because two mutually contradictory criteria can be used to minimize the probability of a Type I error. The first is to choose the model which explains the most variance in the dependent variable. Ordinarily, this will result in choosing the model based on aggregated information. The coefficients of multiple determination in Table 6 are consistent with that expectation. Choosing the aggregate model in order to explain a large proportion of the variance of the dependent variable tells the analyst that each observation lies reasonably close to the plane

Table 6: Aggregated and Disaggregated Coefficients, by State Expenditure Function^a

Function	Constant	Partial Regression Coefficients		Multiple R ²
		Per Capita Federal Aid ^b	Per Capita Income	
<u>Aggregate</u> (N = 16)				
Education	-2047.18	1.05 (.10) ^c	.044 (.001)	.98
Highways	-245.09	-.47 (.16)	.022 (.044)	.90
Welfare	243.38	1.04 (.34)	-.004 (.003) ^d	.94
<u>Disaggregate</u> (N = 768)				
Education	13.00	3.82 (.14)	.010 (.001)	.64
Highways	18.42	.34 (.02)	.004 (.0006)	.34
Welfare	-4.27	.46 (.02)	.004 (.0003)	.52

Source: Data used to generate these estimates are from U.S. Bureau of the Census. State Government Finances (Washington, D.C.: U.S. Government Printing Office, 1951-1968).

Notes:

- a. State expenditures are per capita expenditures by the state alone; federal aid is not included in the dependent variable.
- b. Federal aid represents a three year moving average of per capita federal aid for the given function received by a state.
- c. Figures in parentheses are standard errors of the associated regression coefficients.
- d. $p(b=0) \leq .05$.

of regression. On balance, his errors in predicting how much states will spend with federal aid will not be too severe.

However, another analyst may also be concerned that his estimate of the value of the aggregate regression coefficient for federal aid lies reasonably close to the true value. Such an analyst will use the second criterion, selecting the model whose coefficients have the smallest standard error of estimate. (These two criteria are not entirely independent. If one model explains all of the variance, there will be no standard error associated with any regression coefficient.) Ordinarily, this choice will be the disaggregated model (Orcutt, Watts, and Edwards, 1968). The disaggregated model has a larger number of observations and its independent variables are less likely to be so highly interrelated as to result in collinearity. Both of these factors will reduce the standard error of regression coefficients. The data in Table 6 are also consistent with this expectation, except for education, where there is little difference. In choosing smaller standard errors, however, explained variance is sacrificed. Because it is not always clear which error is the more tolerable, selecting between the two models is even more difficult. The difficulty is compounded if it is unclear whether the probability of Type I or Type II errors should be minimized. In general,

minimizing the chances of one kind of error increases the risk of the other.

Thus the analyst cannot choose between results drawn from aggregated and disaggregated information. Although his first response might be to choose those results which will minimize the probability of Type I errors, it is not clear how these errors should be minimized. By increasing explained variance, he increases the chance that the estimated regression coefficients, which in this case are of particular substantive interest, are not close to their true values. In minimizing the standard error, the model as a whole becomes less accurate as a predictive device.

In sum, public officials demand aggregated information to reduce their decision costs. Yet reduction in decision costs increases uncertainty about individual variations likely to be of concern to a public official. Moreover, there is no clear way to aggregate information. Different summary techniques produce different conclusions and choosing among alternative conclusions is not easy. However, of the two econometric techniques for data aggregation presented in this chapter, the second may provide a reasonable compromise between the demand for aggregated information and the demand for reasonable certainty that aggregated information does not obscure individual variations.

The compromise requires the public official to assimilate slightly more information than would be necessary from the examination of simple aggregate estimates. In return, he receives more information about sources of individual state variation and about the conditions under which the information is most prone to error. The methodological tool to achieve this compromise uses both aggregated and disaggregated information to break the coefficients estimated by regression equation (60) into four basic components.

The first component is a simple average of coefficients estimated for each individual state. The second and third measure the differential responsiveness of each state to nation-wide changes in federal aid and income, respectively. The fourth component suggests the conditions under which estimates of individual state variations are likely to be most error-prone. Thus, instead of three pieces of information, the compromise requires four pieces of information for each federal aid category, or twelve pieces of information. What recommends the compromise is that the second, third and fourth components yield information about individual variations.

What is sought is an expression of the aggregate relation (60):

$$y_t = a + b x_t + c z_t + e_t$$

where

$$y_t = \begin{bmatrix} y_{it} \\ 1 \end{bmatrix}$$

$$x_t = \begin{bmatrix} x_{it} \\ 1 \end{bmatrix}$$

$$z_t = \begin{bmatrix} z_{it} \\ 1 \end{bmatrix}$$

in terms of its more detailed parts. To this end, it is necessary to define three new regression equations for each of the $i = 1, \dots, 48$ states. The first of these new equations is simply the regression of per capita state spending for a given function on per capita federal aid received by the state for that function and on per capita income in the state, where each variable is measured over $t = 1, \dots, 16$ years:

$$y_{it} = a_i + b_i x_{it} + c_i z_{it} + e_{it} \quad (61)$$

This is an individual state regression equation which corresponds to the aggregate equation (60). It is possible to estimate 48 of these equations, for each of the three functional areas, yielding, of course, the massive and unwieldy amount of information a decision-maker would rather avoid. The second new equation for each state is the regression of an independent variable in (61), per capita federal aid received by a particular state, on the two independent variables in (60), which are the national totals for per capita federal aid and per capita income:

$$x_{it} = A_i + B_i x_t + C_i z_t + v_{it} \quad (62)$$

There will be one such equation for each of $i = 1, \dots, 48$ states, where each variable is measured over $t = 1, \dots, 16$ years, and where A_i is the new constant, B_i and C_i are the new partial regression coefficients, and v_{it} is the new error term. Finally, a similar equation regressing the other independent variable in (61), per capita income in a given state, on the two independent variables in (60), must be defined for each state, where each variable is measured over 16 years:

$$z_{it} = D_i + E_i x_t + F_i z_t + w_{it} \quad (63)$$

Analogous to (62), D_i is a constant; E_i and F_i are partial regression coefficients, and w_{it} is the error term associated with equation (63).

The purpose of equations (62) and (63) is to show how changes in two aggregate variables, national per capita federal aid for a given function and national per capita income, affect federal aid and income in a given state. Thus, they relate individual behavior to changes in the values of aggregates. For example, in equation (62), the value of B_i represents how much per capita federal aid received by a given state would change if the national aid budget increased by one dollar per capita, assuming that national income remained constant. The value of C_i measures the change in federal aid received by a state as a

function of a one dollar increase in national per capita income, assuming other variables are constant. In equation (63), E_i and F_i estimate change in state per capita income as a function of a one-dollar increase in national per capita federal aid and national per capita income, respectively. Of particular concern will be the values which relate the state-level variables to their corresponding national aggregates. Thus, for example, an estimated value of B_i for a given state greater than unity indicates that that state receives more than a one-dollar increment in per capita federal aid for each per capita dollar increment in nation-wide federal aid. In a word, that state increases its share of federal aid when federal aid is expanding.

The purpose of this explication is to use this information about individual state responses to changes in the national context in constructing an aggregate. This can be done by substituting (62) and (63) into (61) and summing the resulting expression over all 48 states.

First, (62) and (63) are substituted into (61):

$$\begin{aligned}
 y_{it} &= a_i + b_i x_{it} + c_i z_{it} + e_{it} \\
 &= a_i + b_i (A_i + B_i x_t + C_i z_t + v_{it}) \\
 &\quad + c_i (D_i + E_i x_t + F_i z_t + w_{it}) + e_{it} . \quad (64)
 \end{aligned}$$

Then, after expanding, rearranging and summing each term in (64) over $i = 1, \dots, 48$ states, a new aggregate expression is derived which is consistent with the original

aggregate (60) but which incorporates information from the individual state equations (61), (62) and (63):

$$\begin{aligned}
 y_t &= \sum_i y_{it} \\
 &= \sum_i a_i + \sum_i b_i A_i + \sum_i b_i B_i x_t + \sum_i b_i C_i z_t + \sum_i b_i v_{it} \\
 &\quad + \sum_i c_i D_i + \sum_i c_i E_i x_t + \sum_i c_i F_i z_t + \sum_i c_i w_{it} + \sum_i e_{it} \\
 &= (\sum_i a_i + \sum_i b_i A_i + \sum_i c_i D_i) + (\sum_i b_i B_i + \sum_i c_i E_i) x_t \\
 &\quad + (\sum_i b_i C_i + \sum_i c_i F_i) z_t + (\sum_i b_i v_{it} + \sum_i c_i w_{it} + \sum_i e_{it}) \quad (65)
 \end{aligned}$$

Each term in parantheses in the third expression of (65) is equivalent to the coefficients estimated by (60), plus an implied sampling error (ISE). Each can also be re-expressed in simpler terms, shown in the second expression of the following equations:

$$\begin{aligned}
 a &= \sum_i a_i + \sum_i b_i A_i + \sum_i c_i D_i + \text{ISE} \\
 &= \sum_i a_i + n\text{Cov}(b_i A_i) + n\text{Cov}(c_i D_i) + \text{ISE} \quad (66)
 \end{aligned}$$

$$\begin{aligned}
 b &= \sum_i b_i B_i + \sum_i c_i E_i + \text{ISE} \\
 &= \bar{b} + n\text{Cov}(b_i B_i) + n\text{Cov}(c_i E_i) + \text{ISE} \quad (67)
 \end{aligned}$$

$$\begin{aligned}
 c &= \sum_i b_i C_i + \sum_i c_i F_i + \text{ISE} \\
 &= \bar{c} + n\text{Cov}(b_i C_i) + n\text{Cov}(c_i F_i) + \text{ISE} \quad (68)
 \end{aligned}$$

$$e_t = \sum_i e_{it} + \sum_i b_{iv} e_{it} + \sum_i c_{iw} e_{it} \quad (69)$$

Table 7 shows the results of breaking each coefficient estimated using the aggregate information required by equation (60) into components. Each coefficient based on the aggregate information was originally reported in Table 6, and is reproduced in Table 7, line 1. The first of these components, shown in line 2, is a simple, unweighted sum or average of the corresponding individual estimates for each state generated by equation (61). Note additionally that summing or averaging individual state estimates is yet a third way of aggregating information which will, in general, yield results different from the other two ways of aggregating already discussed.

The second component, line 3, is a weighted sum of individual state coefficients estimated by equation (61). Thus, the second component is itself the sum of two other components, lines (3a) and (3b). The weighting factors for each state are estimated by equations (62) and (63). Each weight gives states with behavior most sensitive to changes in the total federal aid budget or to changes in national wealth greater weight in the aggregate. Thus, if states did not respond to changes in the national context, these weights would disappear and line (3) would be zero. Or, if all states responded in the same way to changes in the national context, these weights would disappear: the coefficients estimated by (62) and (63) would be identical

for each state; hence, the covariance terms in lines (3a) and (3b) would disappear because the covariance of a variable with a constant is zero.

The last component is the implied sampling error. Its value suggests the contextual conditions under which confidence in estimates of individual state behavior can be greatest. It will be discussed more fully below. In sum, lines 2, 3a, 3b and 4 in Table 7 simply break out the elements of line 1. Thus, one piece of information to evaluate the effects of federal aid on state spending in three functions has been expanded to four informational items, providing considerably more insight into differential state responses to contextual changes.

Of particular interest in Table 7 are the results for the components of the aggregate estimate of the federal aid coefficient. The largest component of the aggregate estimate is simply the unweighted average of individual state coefficients. But this unweighted average is larger than the aggregate estimate based on aggregated information in the case of education, it is only apparently identical in highways, and is smaller in welfare. It is reasonable that virtually none of the discrepancy is due to differential responsiveness of state income to changes in the national federal aid budgets; the values in line 3b are very small. In all three cases, a substantial portion of the discrepancy is accounted for by differential effects

of federal aid on individual states due to changes in the total federal aid budget (line 3a).

In education and welfare, the positive covariance terms indicate that states whose spending increases most with additional federal aid are also those whose receipt of federal aid increases most with increases in the total amount of federal aid available. This suggests that, in education and welfare, aid is most stimulative of state spending in states which get the largest increases in federal receipts relative to population.

The opposite is true in highways, where the negative covariance term suggests that states whose spending decreases most with increases in federal aid--that is, states where federal aid is substituted most for state spending--are also those whose federal receipts increase most with increases in the total amount of federal aid. Thus, in highways, when the total federal aid budget increases, states which receive relatively more of that increase are the states most likely to use that money to replace their own spending.

These findings have implications for the direction of federal aid policy. In highways, for example, states most likely to use federal funds to replace their own funds are rewarded by getting more federal aid when the federal aid to highway budget increases. In education and welfare, states which use federal aid to stimulate their own spending are similarly rewarded. Whether these patterns are considered

Table 7: Components of Aggregate Estimates, by State Expenditure Function

<u>Coefficient of per capita federal aid</u>	<u>Education</u>	<u>Highways</u>	<u>Welfare</u>
(1) ^a Total: estimated aggregate coefficient (b)	1.05	-0.47	1.04
(2) Average of corresponding microcoefficients ($1/n \sum b_i$)	1.43	-0.47	.47
(3) Aggregation information	.12	-.11	.235
(3a) Due to $b_i [nCov(b_i B_i)]$	.10	-.10	.230
(3b) Due to $c_i [nCov(c_i E_i)]$	.02	-.01	.005
(4) Implied sampling error	-.50	.11	.335
<u>Coefficient of per capita income</u>	<u>Education</u>	<u>Highways</u>	<u>Welfare</u>
(1) ^a Total: estimated aggregate coefficient (c)	.044	.022	-.004
(2) Average of corresponding microcoefficients ($1/n \sum c_i$)	.043	.019	.001
(3) Aggregation information	.0005	.0003	-.0021
(3a) Due to $b_i [nCov(b_i C_i)]$	.0010	.0001	-.0020
(3b) Due to $c_i [nCov(c_i F_i)]$	-.0005	.0002	-.0001
(4) Implied sampling error	.0005	.0027	-.0029

Source: Data used to generate these estimates are from U.S. Bureau of the Census, State Government Finances, 1951-1968.

Note:

^a Coefficients reported in line 1 are identical to the aggregate coefficients reported in Table 1.

beneficial or damaging depends, of course, on the purpose of each aid program.

A substantial portion of the discrepancy between the two estimates of the effects of federal aid on state spending is also due to implied sampling error. It can be shown (Irwin, 1975) that the implied sampling error represents the coefficient of the regression of the sum of the error terms from the individual state regressions (61) on the independent aggregate variable, which, in this case, is federal aid. Thus, if α represents the value of the implied sampling error in the estimate of the aggregate effects of federal aid on state spending, that value can be estimated by the following regression:

$$\sum_i (y_{it} - \hat{y}_{it}) = \sum_i e_{it} = \alpha x_t + \gamma_t$$

where γ_t is a disturbance term. The implied sampling error in the aggregate estimate of the effects of federal aid to education on state educational spending is negative. This means that the errors made in predicting individual state behavior are smaller when the federal aid to education budget is increasing. The corresponding implied sampling error is positive for highways and welfare. In those areas, errors increase with increases in the federal aid budget. This kind of information is important. It specifies the conditions under which the analyst can have greater or lesser confidence in his predictions about variations in individual state behavior.

Shown also in Table 7 are the components of the aggregate estimate of the responsiveness of state spending to changes in internal resources available to the state. They do not appear large enough to be of immediate substantive interest. In fact, examination of the individual state equations (61) suggests that this responsiveness varies little from one state to another, within a given program. Consequently, the covariance terms are very small.

In brief, Table 6 shows that conclusions about the effects of federal aid on state spending are in part a function of whether aggregated or disaggregated information is used to study the question. Table 7 indicates how breaking the aggregate estimate into components reveals more information than using aggregated or disaggregated information alone. Specifically, analysis of the components suggests that the extent to which federal aid stimulates or replaces state spending as well as the analyst's confidence in his conclusions depends on whether the individual federal aid budgets are increasing or decreasing. Budgetary strategies may be different in an environment of expanding resources than in an environment of declining resources.

This example began by posing a dilemma facing the policy analyst. Public officials demand aggregated summary information. However, summary information obscures individual variations which are of concern to a public official, thereby reducing certainty. Additionally, there is no

clear way to summarize: variables can be summed and then estimated; estimates for individual units can be made, and then summed or averaged; all information can be used to derive unweighted estimates; or some weighting scheme can be used. Each choice will, ordinarily, give different substantive results, and choosing among them will not be easy.

A technique for breaking the first mentioned summarizing device into components was presented and applied to show that it might represent a compromise between the demand for aggregated information and the demand for reasonable certainty that aggregated information does not obscure systematic deviations. Analysis of the components revealed that states whose officials are most successful in using federal education and welfare monies to stimulate state spending are rewarded with more federal aid for their efforts, when the federal aid budget is expanding. Moreover, state officials who are most successful in using federal highway aid to replace state highway spending are similarly rewarded. Significantly, these findings would not emerge from an analysis of aggregate or disaggregate information alone.

II.4 Summary and Conclusions

The purpose of this chapter has been to summarize a wide variety of techniques for aggregating microinformation. Each entails different assumptions and has different consequences.

To aggregate microinformation before estimating macro-relationships results in a significant loss of information about individuals. Nevertheless, it yields information which is easier to interpret than the information generated by procedures which rely on estimation before aggregation. Moreover, the former strategy has a straightforward application in both longitudinal and cross-sectional models, unlike the latter which is most appropriate only in longitudinal models.

If a decision to aggregate before estimating macro-relations is made, further decisions about alternative ways to group the microdata must then be made on the basis of their effects on the bias and efficiency of macroestimates. Here, the dilemma that must be faced is that there is a trade-off between bias and efficiency (Feige and Watts, 1972:354). Grouping procedures which do not bias regression estimates, such as random grouping and grouping by the independent variable, will ordinarily result in a loss of efficiency. On the other hand, grouping by values of the dependent variable will result in bias but may not reduce efficiency. Under certain circumstances, this procedure may result in only a small biasing effect with little loss of efficiency, and may thus be preferable to other procedures. What those circumstances are, however, remain to be determined.

Another dilemma must be faced in deciding to aggregate before estimating macrorelations. Except for random

grouping, all other grouping procedures will bias the correlation coefficient (as well as any other standardized measure of association such as beta weights). However, grouping by the independent variable will not bias estimates of the regression coefficient, an unstandardized measure. Therefore, on the surface it appears reasonable to conclude that use of standardized measures of association should simply be avoided when aggregate relations are estimated from microinformation. However, reflection suggests that, in policy research, this is an unsatisfactory resolution. Policy research is concerned with comparing the influence of alternative policy variables on some behavioral outcome or with comparing the influence of a variable manipulable by policy with other, nonmanipulable, variables. Unless each of the variables whose influence is being compared are measured on the same scale, such as dollars, standardized measures of relative influence must be relied on. On the one hand, converting sociological variables to a common scale will undoubtedly result in invalid measures and measurement error; on the other, use of unstandardized measures will result in aggregation errors. The only viable resolution of this dilemma is to use both standardized and unstandardized measures of association. The former will be unbiased but will not permit conclusions about relative influence, while the latter will be biased but will permit conclusions about relative influence to be drawn. If bias is constant among all the independent variables whose influence is being

compared, conclusions about relative influence will be sound, although conclusions about the magnitude of influence are likely to be exaggerated.

Finally, the strategy of estimating microrelations before aggregating them may be deemed worth additional complexity because it yields additional information about the magnitude and source of aggregation bias. Use of this strategy necessitates temporal data in most cases. It also necessitates avoiding naive assumptions about the specification of the microrelations and the auxiliary relations. However, avoiding naive assumptions may make use of this approach less useful, because more complex solutions will be necessary. This only exacerbates the difficulties of interpreting results based upon this approach.

III. METHODS OF DISAGGREGATION:

EFFECTS ON BIAS

III.1 Introduction

Demonstrating the effects of alternative methods for aggregating information also defines the effects of disaggregation. The conditions under which aggregation will not bias estimates of correlation and regression coefficients are identical to the conditions under which disaggregation yields unbiased estimates. However, the purpose of disaggregation is to recover unavailable information about individual level behavior, while the purpose of aggregation is to summarize available information about individual level behavior. Consequently, using aggregate information to make inferences about individuals will not simultaneously affect the efficiency of estimation. Moreover, because disaggregation problems arise in the absence of information, these problems are somewhat more complex than those of aggregation. Without information about individuals, there are almost an unlimited number of possible inferences about individual behavior that can be made from a given set of information about aggregates. Only theories about individual behavior can constrain the number of alternative inferences. Yet these theories are often completely or largely untested at the individual level. Had they been tested at the

proper level, the disaggregation problem would not have existed in the first place.

Problems of disaggregation are common in educational research. The cost of gathering data about individual students or schools is very high. The administration of such detailed data collection efforts is also complex. The more units for which data are collected, the greater the problems of comparability. This dilemma is particularly intractable when financial and administrative information about schools and programs within schools is sought. Large quantities of information about individual students and schools are not only difficult and costly to gather; they are also difficult to assimilate and understand. As a result, information about individuals is often available only for small or unique areas. Information is frequently available only for, say, a few cities, as in Crain's study of school desegregation in 10 cities (Crain, 1969). Information may be available only for a few states, as in the Ford Foundation educational finance project (Berke et al., 1973). Or information may be available only for areas of particular interest, as in the studies of large-city school spending (Burkhead, et al., 1967). In a word, information about individual behavior is frequently not representative. Yet policy makers demand information representative of the nation as a whole. In the absence of adequate funding for large-scale

studies of the magnitude of the Equal Educational Opportunity Survey, and because the results of such large-scale studies are frequently not timely enough to meet the needs of policy-makers, aggregate information must frequently be used to make inferences about the behavior of students, schools and even school districts.

This chapter explicates the three most common and useful strategies for disaggregating aggregate information. These three strategies are the method of bounds, ecological regression and correlation, and proper specification of the aggregate relation. Used in conjunction, these three strategies will, under many conditions, yield unbiased estimates of individual behavior. Moreover, two of the strategies are applicable with dichotomous, polytomous, and interval level data, with bivariate and multivariate relations; and with longitudinal and cross-sectional measurement.

III.2 General Approaches

III.2.1 The Method of Bounds

The method of bounds was first suggested by Duncan and Davis (1953). It does not yield exact estimates of individual level behavior from grouped data, but only ranges within which the true estimates must lie. Nor does it yield information about regression coefficients, for it deals only with correlations. Although the method could be adapted to yield estimates of the ranges of

microlevel regression coefficients or their analogies from macrolevel data, ecological regression, another method discussed below, yields a single estimate and is, therefore, more useful. Unlike ecological regression, the method is suitable only for categorical data or for interval level data which have been categorized. Categorizing interval level data only results in a loss of information in addition to the loss already generated from its aggregated nature. Because of these and other defects in the method of bounds which will be discussed subsequently, its primary use is only as an external check on the validity of ecological correlation estimates.

The method is simple and straightforward. Suppose that interest centers on the correlation between race and achievement levels of students enrolled in public schools. Data, however, are available only on the number of whites and non-whites in school districts (or in a representative sample thereof) as well as on the number of students in each of these districts who have achieved above and below the national norm. For each district, a cross-classification of race and achievement is sought but only the marginals are known. Knowing the marginals alone permits only the computation of the between-group or macrolevel correlation. The internal cell values must be known or inferred to permit computation of the individual level correlation. However, the value of the marginals does deter-

Table 1: Achievement by Race, Marginals Known

	White	Non-White	
Above nat'l norm			10,638
Below nat'l norm			5,239
	12,324	3,553	15,877

mine the highest and lowest values that the individual level correlation could have. For instance, Table 1 shows hypothetical marginal frequencies for a single district.

At the minimum, in this school district, no non-whites could have achievement scores over the national norm. If that were true, the remaining cells would have the frequencies shown in Table 2. At the maximum, all the non-whites could have scored above national norms.

Table 2: Achievement by Race: No Non-Whites Above National Norm

	White	Non-White	
Above nat'l norm	10,638	0	10,638
Below nat'l norm	1,686	3,553	5,239
	12,324	3,553	15,877

The cell frequencies would, therefore, be those in Table 3.

Table 3: Achievement by Race: All Non-Whites Above National Norm

	White	Non-White	
Above nat'l norm	7,085	3,553	10,638
Below nat'l norm	5,239	0	5,239
	12,324	3,553	15,877

For this hypothetical school district, the lowest correlation (a fourfold point correlation, or phi) that could occur, given the marginals, is -0.38 computed from Table 3. The highest is 0.77 , computed from Table 2. Replicating this process successively for all school districts will produce similar minimum and maximum correlations. Choosing the least minimum and greatest maximum of the correlations computed in this fashion provides the absolute upper and lower bounds within which the individual level correlation must lie. (These are not confidence intervals; rather, it is a certainty that the individual correlation lies within these bounds.)

The defects of this method should be clear. First, it can produce such wide bounds that they will be of little practical use. However, the bounds are likely to be smaller when smaller units or groups are used. The bounds produced when school data are used to make inferences

about students are likely to be smaller than when school district data are used. Of course, the gain in utility comes at the expense of more disaggregated data. Second, the method becomes more awkward to employ with polytomous as compared to dichotomous data. In a 3×2 table, given the marginals, two independent internal cell frequencies must be specified before the remaining cell frequencies can be determined. In a 3×3 table, four cell frequencies must be determined. In an $r \times c$ table, $(r - 1) \times (c - 1)$ cell frequencies must be determined. Given the marginals for a particular unit, which of the $(r - 1) \times (c - 1)$ cell frequencies yields the maximum and minimum correlation for that unit becomes less obvious and must be determined empirically. Moreover, general $r \times c$ tables have no measure of association like the fourfold point correlation, which is identical to the Pearson product-moment correlation in the 2×2 case alone.

Thus the method is most useful with fourfold tables. The method is also applicable with time series data. For example, in the absence of individual level data, to find whether students' first grade achievement levels relative to national norms bear any relation to their relative 12th grade achievement levels, known school district marginal frequencies can be employed to estimate the individual level correlation by the method of bounds. (See Table 4.) Finally, the method can be applied in a straightforward

Table 4: Twelfth Grade Achievement by First Grade Achievement: Hypothetical School District

Twelfth Grade (1972) Above Below	First Grade (1950)	
	Above	Below

manner to partial relations. For instance, the method of bounds could be applied separately to Table 4 for predominantly Black and predominantly white school districts.

In contrast to the method of bounds, the method of ecological regression and correlation has few of its defects and all of its advantages. Consequently, the method of bounds is used simply to check the other's validity.

III.2.2 Ecological Regression and Correlation

The linear regression of an aggregate dependent variable on one or more aggregate independent variables can be used to make inferences about individual level regression and correlation coefficients under certain special circumstances. This method, first proposed by Goodman (1953) and later extended (Goodman, 1959), is called ecological regression. It is equally useful with 2×2 and general $r \times c$ tables, with bivariate as well as multivariate relations, and with cross-section as well as longitudinal data. Perhaps most importantly, although linear ecological regression is appropriate only if certain

restrictive conditions are met, many non-linear extensions are possible to help circumvent this restriction. In this section, only linear ecological regression will be discussed. Nonlinear extensions will be considered in subsequent sections of this chapter.

The logic of linear ecological regression is easiest to grasp using a 2 x 2 table. Consider the relation between achievement and race shown in Table 5. The following terms can be defined from the frequencies labeled

Table 5: Achievement by Race: Cell and Marginal Frequencies

	White	Non-White	
Above nat'l norm	a	b	a + b
Below nat'l norm	c	d	c + d
	a + c	b + d	N

there:

$Y = (a + b)/N$ = proportion of (school district or state) population which scores above the national median.

$X = (a + c)/N$ = proportion of (school district or state) population which is white.

$p = a/(a + c)$ = proportion of whites (within school district or state) who score above national median.

$r = b/(b + d)$ = proportion of non-whites (within school district or state) who score above the national median.

From these definitions, it follows that Y can be written by:

$$Y = p(X) + r(1 - X) \quad (1)$$

Collecting terms, Y can be rewritten:

$$Y = r + (p - r)X \quad (2)$$

Equation (2) is a simple bivariate linear relation of the form

$$Y = a + bX \quad (3)$$

If the proportions p and r are constant across all ecological units, such as states or school districts, then there will be an exact relation between Y and X for all of the units. A straight line fitted to the observed macrolevel data points can be used to determine $r = a$ and $p = b + a$, where r and p are the missing microlevel internal cell proportions.

Of course, for p and r to be invariant within each ecological unit would be highly unusual. However, regression procedures can still be used if the expected values of p and $(p-r)$ are constant. In this situation, linear regression is still appropriate:

$$E(Y) = E[r(X)] + E[(p-r)(X)] \quad (4)$$

which is of the form:

$$y = \alpha + \beta X$$

where

$$\alpha = E[r(X)] \text{ and } \beta = [E(p-r)(X)] .$$

Unbiased estimates of α and β are obtained by ordinary least squares procedures:

$$y = a + bX \quad (5)$$

Having estimated p and r from a and b , the individual level correlation can easily be computed from the four-fold point correlation. This correlation uses the cell frequencies of Table 5 which can be derived from p , r and the known marginal frequencies. Specifically, the frequency estimates are given by pNX , $rN(1 - X)$, $(1 - p)NX$, and $(1 - r)N(1 - X)$, where NX is the total number of non-whites. However, the fourfold point correlation is analogous to the product-moment correlation only in the 2×2 table. Goodman suggests a more general approach which has the added virtue of simplicity. Specifically,

$$r = b \sqrt{X(1 - X)/\hat{Y}(1 - \hat{Y})} \quad (6)$$

If ecological regression is applicable, $\hat{Y}$ will be close to Y . Hence, $\hat{Y}$ can be replaced by Y , so that:

$$r = b \sqrt{X(1 - X)/Y(1 - Y)} \quad (7)$$

Goodman also points out that if the residual variance of Y , $\sigma_{y/x}^2$ is constant for all values of X , a and b will be both unbiased and the most efficient estimates of α and β . Moreover, if, in addition to homoscedasticity, the distribution of Y for a given X is normal, then it is also possible to obtain confidence intervals for β , $\hat{a}$, $\hat{Y}$ and r .

The most important criterion is that a and b be unbiased estimates, so that p and r can be derived. This requires that

any variation in p and r across ecological units is purely random. Specifically, in order to use linear ecological regression, the proportion of non-whites in an ecological unit who score below the national median must be independent of the proportion of non-whites in that unit. Likewise, the proportion of whites who score below the national median in an ecological unit must be independent of the proportion of non-whites in that unit. A non-white attending school in a predominantly non-white district must be assumed to score as well (or poorly) on an achievement test as his hypothetical twin would in a predominantly white school district. The same applies to hypothetical white twins.

Clearly, the validity of these assumptions cannot be tested in the absence of individual level data. If the individual level data were available, ecological regression would not be necessary in the first place. Nevertheless, Goodman provides a number of tests for the validity

of these assumptions which rely only on the aggregate or ecological data. Meeting these tests, however, is only a necessary but not a sufficient condition for the validity of these assumptions. However, with the exception of the linearity test, each of these tests is equally applicable for the nonlinear extensions discussed below.

First of all, Goodman points out (1959:613) that if assumptions for linear ecological regression appear invalid for the whole set of ecological units, the method may still be applicable for separate subdivision estimates of individual level regressions and correlations.

Second, a necessary but not sufficient condition for linear ecological regression is that the scatter diagram of Y and X be linear (Stokes, 1969). Alternatively, the scatter diagram for a subset of ecological units should suggest linearity. If the scatter diagram does suggest linearity, it is then possible that $E[p(X)]$ and $E[r(X)]$ are constant. However, the variances of p and r may still be quite large, making estimates of the individual level r quite poor.

Third, since $E[p(X)]$ and $E[r(X)]$ must lie between 0 and 1, it is desirable that their estimators, p and r , also lie between 0 and 1. It is possible to obtain values of p and r which lie outside of this range, even if the assumptions for linear ecological regression are met. Yet failure to obtain such values should still lead to suspicion that the assumptions may not be valid. Goodman

(1953) suggests that if the assumptions are satisfied, and if a value such as $r = -.05$ still results, its true value can be assumed to be zero, and a new least squares estimate for p using the equation $E(Y) = pX$ should be determined.

A fourth criterion that an ecological regression should meet is that the estimated proportion, $\hat{Y} = \hat{a} + \hat{b}X = pX + r(1 - X)$, should be fairly close to the known proportion Y , where $\hat{Y} = \sum_i \hat{Y}_i$. In the example presented above, this means that $\sum_i \hat{Y}_i$ should be close to $Y = .50$, since 50% of all students nationally must score below the national median. This criterion will not work if $Y = \bar{y}$ and $X = \bar{x}$, because in this case $\hat{Y} = \hat{a} + \hat{b}\bar{x} = \bar{y} = Y$, even if the assumptions are not met.

Fifthly, the estimated variance $\sigma_{Y/X}^2 = \sigma_{\hat{Y}}^2 = \sigma_{\bar{Y}}^2 + (X - \bar{x})^2(\sigma_b^2)$ should be fairly small, where

$$\sigma_{\bar{Y}}^2 = \sigma_{Y/X}^2 / (1/n) \text{ and } \sigma_b^2 = \sigma_{Y/X}^2 / n\sigma_x^2$$

This criterion should be used in conjunction with the observed difference between Y and $\hat{Y}$ to determine whether these two values are sufficiently close.

The individual level correlation coefficient estimated by (6) should, of course, lie within the bounds estimated by the method discussed in the preceding section.

Finally, Goodman suggests that if it is unreasonable to assume that the column-based cell proportions p and r

implied by Table 5 are not constant, an alternative, more viable assumption might be to assume that the row-based cell proportions are constant. In other words, it could be assumed that the proportion of non-whites who score below the national median and the proportion z of non-whites who score above the national median is roughly constant across ecological units. In this case, referring back to Table 5, $q = a/(a + c)$ and $z = c/(a + c)$. Of course, only their expected values, $E[q(Y)]$ and $E[z(Y)]$ need be constant. Ecological data measuring Y and X can then be used to estimate q , z and r , the correlation. The following equation defines X :

$$\begin{aligned} X &= qY + z(1 - Y) \\ &= z + (q - z)Y \end{aligned} \quad (8)$$

which is of the form:

$$X = c + dY \quad (9)$$

where $c = z$ and $d = (q - z)$. If the assumptions for linear ecological regression are met, then ordinary least squares applied to (9) yields unbiased estimates of the unknown individual cell proportions, $E[q(Y)]$ and $E[z(Y)]$.

This approach has been extended even further by Shively (1969). He argues that use of the "row model" and "column model" in conjunction can detect the direction of bias in estimates of the unknown cell proportions. According to Shively, if data are grouped on the row

scores (Y), and if there is a positive ecological relation between X and Y, the value of p will increase with increasing values of X and a row model estimate for p will exceed its actual value. When the marginals X and Y are negatively related, p will decline as X increases, and the estimate for p will be less than its true value.

The utility of this conclusion would be clear, if it were valid. Meckstroth (1974) strongly criticizes its validity, however. He points out that Shively's conclusion presumes knowing how the data are grouped--whether by Y or by X, because the conclusions are valid only for data grouped by the dependent variable. If there is enough information to accept an assumption about how the data are grouped, "there will also be sufficient grounds to select one set of estimates (e.g., from the column model) and be assured that these are nearly accurate" (Meckstroth, 1974: 58).

A more propitious means of relaxing the restrictiveness of the assumptions for linear ecological regression is to use nonlinear extensions. While these will be discussed subsequently, it is also important to understand the widespread applicability of both linear and nonlinear ecological regression. Its extension, for example, from a 2×2 to an $r \times c$ array is straightforward.

Table 6: $r \times c$ Array-Marginal and Cell Proportions

P_{11}	P_{12}	P_{13}	$P_{14} = P_{1c}$	$Y_{i=1}$
P_{21}				$Y_{i=2}$
$P_{31} = P_{r1}$				$Y_{i=3}$
$\underbrace{X_{j=1} \quad X_{j=2} \quad X_{j=3} \quad X_{j=4}}_{j=1, \dots, c}$				$i=1, \dots, r$

In Table 6, the marginal proportion for the i^{th} row, Y_i , is related to the cell proportions by the identity

$$Y_i = P_{i1}X_1 + P_{i2}X_2 + P_{i3}X_3 + P_{ic}X_c \quad (10)$$

where

$$X_c = 1 - X_1 - X_2 - X_3 = 1 - X_1 - X_2 - X_{c-1}$$

Therefore, (10) can be rewritten:

$$\begin{aligned}
 Y_i &= P_{i1}X_1 + P_{i2}X_2 + P_{i3}X_3 + P_{ic}(1 - X_1 - X_2 - X_{c-1}) \\
 &= P_{i1}X_1 + P_{i2}X_2 + P_{i3}X_3 + P_{ic} - P_{ic}X_1 - P_{ic}X_2 - P_{ic}X_{c-1} \\
 &= P_{ic} + (P_{i1} - P_{ic})X_1 + (P_{i2} - P_{ic})X_2 + \\
 &\quad (P_{i3} - P_{ic})X_{c-1}
 \end{aligned}$$

More generally, the marginal proportion for the i^{th} row of any $r \times c$ array can be written:

$$Y_i = P_{ic} + (P_{i1} - P_{ic}) X_1 + (P_{i2} - P_{ic}) X_2 + \dots + (P_{ij} - P_{ic}) X_j + \dots + (P_{i,c-1} - P_{ic}) X_{c-1} \quad (11)$$

which is of the form required for linear ecological regression:

$$y_i = a + b_1 X_1 + b_2 X_2 + \dots + b_{c-1} X_{c-1},$$

assuming that the P_{ij} are roughly constant over all ecological units. Successive application of (11) to each of $r - 1$ rows in an $r \times c$ matrix allows each cell proportion to be recovered.

Many of the ideas presented above can also be applied to interval level variables when information about individual level relations is desired but unavailable. For example, suppose that knowledge about the relation between median family income and expenditures per pupil for schools is needed, but data for these variables are available only for school districts (or states). If there is a linear relationship between these variables, and if the regression of per pupil expenditures on median family income (i.e., $E(y/x) = A + BX$) is relatively constant in all school districts (or states), so that A and B are constant, the ecological data can be used to estimate the school-level relationship (Goodman, 1959:620). The appropriate estimate

of the school-level Pearsonian correlation from the ecological data is obtained by multiplying $\hat{b}$, the ecological regression, by a ratio V . V is the population standard deviation (or variance) of the independent variable divided by the population standard deviation (or variance) of the dependent variable. The population of interest in the example consists of schools. Thus, the individual level correlation is $\hat{b} \sqrt{\sigma^2(X)/\sigma^2(Y)}$. The school-wide variance of income is

$$\sigma^2(X) = [\sum_{i,j} w_j (X_{ij} - \bar{X}_j)^2 + \sum_j w_j (\bar{X}_j - \bar{X})^2]$$

and $j = 1, \dots, J$ ecological units, $i = 1, \dots, I_j$ schools in the j^{th} unit, and w_j is the relative population size (or number of schools) in the j^{th} unit. Thus $\sigma^2(X)$ is the total variance of median family income for schools. Its components are the within-unit and between-unit variances respectively. Likewise, the school-wide per pupil expenditure variance is:

$$\sigma^2(Y) = [\sum_{i,j} w_j (Y_{ij} - \bar{Y}_j)^2 + \sum_j w_j (\bar{Y}_j - \bar{Y})^2].$$

In contrast, the usual ecological correlation is

$$\hat{b} \sqrt{BS(\bar{x})/BS(\bar{y})},$$

where $BS(\bar{x}) = \sum_j (X_j - \bar{X})^2$ and $BS(\bar{y}) = \sum_j (Y_j - \bar{Y})^2$. For reasons already discussed, the ecological correlation is likely to be larger than the individual level correlation.

Estimation of the individual level correlation does not require knowledge of the entire joint distribution of the independent and dependent variables for each individual within each ecological unit. However, it does require knowledge of either the within or total variance of each variable. Reliable estimates of within-unit variances are unlikely to be available; if the estimates were available, the same data would also permit direct estimation of within unit regression coefficients from individual level data. Estimates of individual level regressions from ecological data, however, need not be sacrificed, because total variance estimates may be obtainable from old but already available individual data, such as the Equal Educational Opportunity Survey and the Decennial Census. While mean values of these data have undoubtedly changed, the variances are probably much more stable. Thus, even though the Survey is dated and the Census material may also be dated, there is no reason to believe that variances have changed. More current variance estimates may also be obtainable from the Current Population Survey. Variance estimates from these sources could be used in other, contemporary, ecological studies with no increment in cost.

Ecological regression with quantitative variables is appropriate only if the expected value of the regression coefficient is constant within each ecological unit. The indirect tests of this assumption suggested above are

weak because the test criteria may not be fulfilled even if this assumption is valid. Nevertheless, findings consistent with these tests increase confidence in the validity of the assumptions. Of the six tests, two are inappropriate if the ecological data are quantitative, and one requires slight modification. First, the method of bounds is generally inappropriate with interval-level data. Second, with quantitative ecological variables the coefficients are usually free to take on any value, unless the variables are proportions. The regressions, unlike cell coefficients for such variables will not lie between 0 and 1. Third, substitution of the known population mean, $\bar{X}$, into the ecological regression $\hat{y} = \hat{a} + \hat{b}X$, should yield a predicted $\hat{y}$ fairly close to the known $\bar{y}$.

The methods discussed above are also applicable to time-series studies. Subsystem change from time 1 to time 2 can be measured with ecological data, for example, if the assumptions of ecological regression are met (Meckstroth, 1970). If data on the progress of desegregation in schools are unavailable or too costly to collect, school district level data on the number of Black and white students measured at two points in time might be an adequate substitute. (See Table 7.) If it can be assumed that this temporal relationship is linear and that the cell proportions, $p = a/(a + c)$ and $r = (b/(b + d))$ are roughly constant over school districts, then the

Table 7: Number of Black and White Students at Two Points in Time; Marginal Frequencies for School Districts

	Time 1		
Time 2	No. of Black Students	No. of White Students	
No. of Black Students	a	b	a+b
No. of White Students	c	d	c+d
	a+c	b+d	

regression $E(y/x) = A + BX$, where $E(y/x)$ is the expected proportion of Blacks at time 2 given the proportion at time 1, $A = \hat{r}$, and $B = (\hat{p} - \hat{r})$, gives unbiased estimates of the unknown cell proportions p and r .

Where the data are quantitative rather than qualitative, the extension to time-series is equally straightforward. Ecological variables simply measure values for serial units of time instead of values for areal units measured at the same time. If the assumptions of ecological regression are met, then its use is warranted. There are additional problems which are peculiar to any kind of least squares regression model using temporal data. Autocorrelation is a particularly intractable problem, but it is not unique to ecological regression and will not be discussed herein.

In conclusion, simple linear ecological regression yields unbiased estimates of individual level correlation and regression coefficients whenever (a) the relation

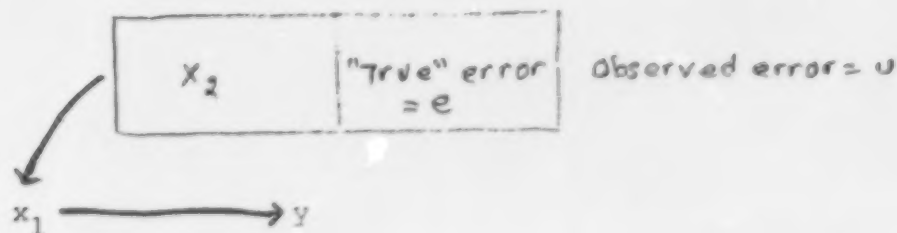
between the variables of interest is linear; (b) the expected values of cell proportions (or regression coefficients) are constant across all ecological units; and (c) the ecological regression is properly specified. If these assumptions are not met for the whole set of ecological units, they may still be met for smaller subsets of those units. Ecological regression can then be applied to each subset to yield unbiased estimates of individual relations. Unfortunately, there is no way to test the third assumption if only aggregate data are available. Moreover, the second assumption is also untestable in the absence of individual level information. However, there are a number of criteria which the results of an ecological regression should meet. Meeting these criteria increases confidence that the assumption are, in fact, valid. However, the criteria constitute only necessary but not sufficient conditions for the validity of the assumptions. Finally, if the assumptions are valid, ecological regression can be used with qualitative as well as with quantitative variables and with cross-section as well as longitudinal data. The following section indicates that ecological regression is also appropriate with multivariate linear models. The use of such models is frequently necessary if ecological regressions are to be properly specified.

III.2.3 Eliminating Bias in the MacroRelation: Multi-variate Linear Ecological Regression

The preceding discussion of ecological regression has implicitly assumed that two variables completely specify the behavior of interest. In other words, it has assumed that a bivariate macrorelation will yield unbiased

estimates of individual level relations if (a) the regression coefficients within units are roughly constant and (b) if there is no specification bias in the macrorelation itself. Condition (b) means that the error term in an ecological regression is uncorrelated with the independent variable in the regression. If such a correlation exists, then the error term includes an unmeasured factor which systematically influences both the measured independent and dependent variables. The omitted factor is called a "confounding" variable (Blalock, 1964). Such a confounding variable is portrayed diagrammatically in Figure 1. If y and x_1 are ecological variables whose regression is

Figure 1: Confounding Variable-- x_2



intended to estimate an individual-level relation, even if condition (a) is met, failure to meet condition (b) will produce a biased estimate of the individual level relationship. Bias is produced because the failure to include x_2 means that the error term u is not independent of the included independent variable x_1 .

How omission of a confounding variable produces specification bias is easily shown. Suppose the true model, implicit in Figure 1, is:

$$y = B_1X_1 + B_2X_2 + e \quad (12)$$

where each variable, for convenience, is measured about its mean; $E(e) = 0$; and $\text{COV}(e, X_1) = 0$. The unbiased estimate of B_1 is $b_{y1.2}$; the estimator for B_2 is $b_{y2.1}$. The misspecified model omits X_2 :

$$y = B'_1X_1 + U, \quad (13)$$

where $U = B_2X_2 + e$. The ordinary least squares estimator of B'_1 is

$$E(b_{y1}) = E[\Sigma(X_1, Y) / \Sigma X_1^2]. \quad (14)$$

Substituting (13) into (14) shows that B'_1 is related to B_1 by the following equation:*

*The derivation is straightforward:

$$\begin{aligned} B'_1 &= E \left[\frac{\Sigma(X_1, Y)}{\Sigma X_1^2} \right] \\ &= E \left[\frac{\Sigma X_1 (B_1X_1 + B_2X_2 + e)}{\Sigma X_1^2} \right] \\ &= E \left[\frac{B_1 \Sigma X_1^2 + B_2 \Sigma X_1 X_2 + \Sigma X_1 e}{\Sigma X_1^2} \right] \\ &= B_1 + B_2 E \left[\frac{\Sigma X_1 X_2}{\Sigma X_1^2} \right] \quad \text{since } E(\Sigma X_1 e) = 0 \end{aligned}$$

$$B'_1 = B_1 + B_2 \cdot E \left[\frac{\sum X_1 X_2}{\sum X_1^2} \right] \quad (15)$$

Equation (15) can be rewritten in terms of its estimators:

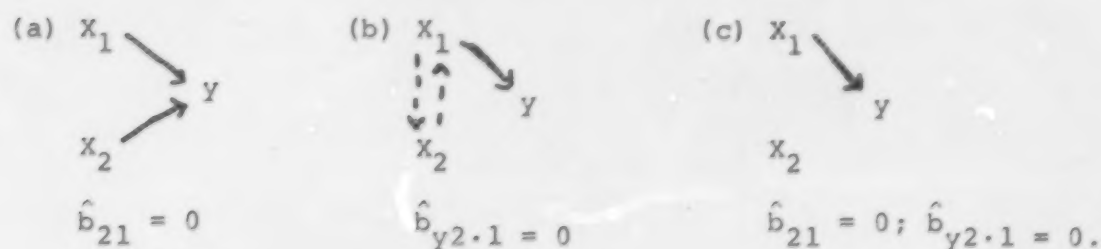
$$\hat{b}_{y1} = \hat{b}_{y1.2} + \hat{b}_{y2.1} \hat{b}_{21} \quad (16)$$

The amount of specification error is the difference between the biased estimator of B_1 , or $\hat{b}_{y1}$, and the unbiased estimator, $\hat{b}_{y1.2}$:

$$\hat{b}_{y1} - \hat{b}_{y1.2} = \hat{b}_{y2.1} \hat{b}_{21} \quad (17)$$

Specification error will disappear if one or both terms on the right hand side of (17) are zero. If X_1 has no direct effects on X_2 , $\hat{b}_{y1}$ will be an unbiased estimator of B_1 . (See Figure 2a.) Alternatively, $\hat{b}_{y1}$ will also be unbiased if X_2 has no independent effects on y . (See Figure 2b.) Finally, $\hat{b}_{y1}$ will be unbiased if both of these conditions are met. (See Figure 2c.)

Figure 2: Conditions for Disappearance of Specification Error



These conditions are unlikely to be met, particularly with ecological data. First of all, the disappearance of specification error requires no collinearity among the independent variables, but ecological data are especially subject to collinearity (Hanushek, Jackson and Kain, 1974). For instance, one of the substantive examples discussed previously dealt with the effects of race on achievement. Yet it is well known that areas which have high proportions of Blacks are also likely to be low income areas, and that children from low income families tend to score more poorly on achievement tests than children from high income families. Family income was a variable omitted from the previous example. Its omission alone insures that $\hat{b}_{y1}$ is a biased estimate of the individual level regression coefficient measuring the effects of race on achievement, even if the within-unit regression coefficients are invariant. Because neither $\hat{b}_{21}$ nor $\hat{b}_{y2.1}$ are likely to equal zero, omitting income will mean that $\hat{b}_{y1}$ is biased.

The direction of bias, though not its magnitude, can frequently be inferred (Lichtman, 1974). It is well known, for example, that high income areas are likely to have lower proportions of Blacks. Consequently, $\hat{b}_{21}$ will be negative. It is also known that areas where family income is higher are likely to have public school students with higher achievement scores. Moreover, this effect is direct,

because family background has a direct effect on student achievement. Consequently $b_{y2.1}$ is probably positive. Specification error is therefore negative; b_{y1} is likely to be smaller than its unbiased value $b_{y1.2}$.

In short, even if the assumptions for ecological regression are met, so that relations are constant within ecological areas, use of an incompletely specified ecological regression will lead to biased estimates of the individual-level relations. Possibly confounding variables must therefore be explicitly introduced into the ecological regression to avoid this source of bias. When the ecological variables are quantitative, adding another variable to the regression equation is straightforward, requiring the use of multiple rather than simple regression. When the data are qualitative, Goodman's method can simply be applied to successive rxc crosstabulations embedded within values of a third variable.

III.2.4 Unique Estimates in the Macrorelation: Non-Linear Ecological Regression

The ecological regression models discussed so far are linear. In other words, the expected values of the cell proportions are assumed to be independent of the value of the marginal proportions. This assumption implies that the cell proportions will be invariant regardless of the value of the marginal proportions.

More specifically, nonlinear models are appropriate when individual behavior varies according to the context in which it occurs. Under certain circumstances, however, the use of aggregate data makes it impossible to separate individual from contextual effects. Yet obtaining unique estimates of the impact of contextual properties on individual behavior is crucial because contextual properties are manipulable while individual properties are not. It is therefore important to specify those contextual properties that impede unique estimation from aggregate data. If those properties are operative, using aggregate data to estimate individual behavior considerably diminishes the utility of aggregate data for educational policy research, even if the aggregate estimates are otherwise unbiased.

Whether unique estimation is possible depends on the type of contextual property that is operative in any given instance. In particular, the source and impact of

contextual variables determines whether they are amenable to aggregate level analysis. The source of a contextual variable specifies how it is related to individual level independent variables. Its impact specifies how it influences the individual level dependent variable.

There are three sources of contextual effects. First, a contextual variable can be a function of an independent variable included in an individual level relation. For example, $\bar{X}_2$ defines the socioeconomic status of those with whom a student attends school; $\bar{X}_2$ is a function of X_2 , the status of individual students. Second, a contextual variable can be a function of the dependent variable included in an individual level relation. The achievement level of a student's peers, $\bar{Y}$, is such a contextual variable. In fact, school policies based on tracking presuppose that the achievement of peers has an independent impact on the achievement of individual students. Finally, the definition of many contextual variables is unrelated to any of the variables included in the individual level relation. Variables like the quality of the teachers and the school's facilities typify this type of contextual variable, denoted by $\bar{Z}$.

Contextual variables can also affect behavior in two ways. First, they can affect behavior directly. Thus, the socioeconomic status of those with whom a student attends school could have a direct effect on his

own achievement. Contextual variables can also have interactive effects. For example, attending school with students of higher status can reinforce the effects of a student's own race or status on his ability. If a high status student has high achievement, that impact may be magnified if the student attends school with other high status students. Alternatively, the impact may be lessened if he attends school with lower status students. With regard to school quality, the impact of low status may be reduced if low status students attend schools with good facilities and high calibre teachers. In fact mitigating the effects of race and status by improving quality was an early goal of Title I of the Elementary and Secondary Education Act of 1965 (McLaughlin, 1975: 34).

Aggregate estimates of the independent impact of contextual variables, however, are often deficient. Consider first contextual variables that arise from independent variables included in the individual level relation, such as the socioeconomic status of a student's peers. Coleman (1966: 304-310) showed clearly that low status students benefit from daily contact in school with higher status students. As the status of their peers increases, low status students are more likely to score higher on standardized verbal achievement tests. At the individual level, peer effects like these could operate in one of two ways. First, they could affect achievement directly:

$$Y = b_1X_1 + b_2X_2 + b_3\bar{X}_2 + e, \quad (18)$$

where each variable was previously defined but the regression coefficients are now unstandardized rather than standardized.

Peer effects could also interact multiplicatively with race. Attending school with high status peers may mitigate the effects of race on achievement. In other words, the achievement differences between blacks and whites in a high status school may be less than the differences in a low status school. Such a model can be formalized by a regression:

$$Y = b_1X_1 + b_2X_2 + e, \quad (19)$$

where $b_1 = (b_3 + b_4\bar{X}_2)$, so that the resultant equation is:

$$Y = b_3X_1 + b_4X_1\bar{X}_2 + b_2X_2 + e. \quad (20)$$

Peer effects could also interact with an individual student's socioeconomic status. In other words, a student's own status may have less impact on his achievement when he attends school with high status peers than when he attends with low status peers. In this case, b_2 in equation (19) is a function of the mean status of those with whom a student attends school:

$$b_2 = (b_3 + b_4\bar{X}_2).$$

By substitution, the appropriate model becomes:

$$Y = b_1X_1 = b_3X_2 + b_4X_2\bar{X}_2 + e. \quad (21)$$

A significant value for b_4 would indicate that a student's own status has differential effects on his achievement, depending on the status of his peers.

Thus, the socioeconomic background of peers constitutes a contextual variable that could affect a student's achievement in three different ways, represented by equations (18), (20) and (21). Each individual level model has a corresponding aggregate model whose variables are measured on schools to make inferences about students. School means therefore replace individual values.

If the effects of contextual variables derived from independent variables included in the individual level relation are direct, the corresponding aggregate model makes it impossible to separate the effects of contextual from individual variables. For example, equation (18) uses data on students to isolate the effects of individual and peer socioeconomic status. Reliance on aggregate data makes it impossible to estimate the unique contribution of those variables, as equation (22) illustrates:

$$\begin{aligned}\bar{Y} &= b_1\bar{X}_1 + b_2\bar{X}_2 + b_3\bar{X}_2 + e \\ &= b_1\bar{X}_1 + (b_2 + b_3)\bar{X}_2 + e.\end{aligned}\tag{22}$$

In this equation, as well as in equation (18), b_2 represents the coefficient of individual status effects, while b_3 represents the coefficient of contextual status effects. In equation (18), however, each coefficient can be uniquely estimated; equation (18) has three unknown and three known values. Equation (22), in contrast, has only two known values. A unique estimate of contextual effects, represented

by b_3 , cannot be obtained unless some a priori value is assumed for b_2 , the coefficient of individual effects. Until more is known about the impact of status on achievement to warrant any such assumptions, all that can be estimated is the total effect of both peer and individual status on achievement.

Thus, if contextual effects derived from an individual level variable included in the individual level equation are direct, unique estimates of their effects cannot be obtained from aggregate data alone. Bereft of unique estimates, an investigator cannot distinguish between a model that includes only individual level effects and one which includes both individual and contextual effects. At the individual level, a model that includes only the effect of a student's own race and status is:

$$Y = b_1X_1 + b_2X_2 + e. \quad (23)$$

In the absence of data on students, equation (23) becomes:

$$\bar{Y} = b_1\bar{X}_1 + b_2\bar{X}_2 + e. \quad (24)$$

Equation (24) is indistinguishable from equation (23), even though (24) omits contextual effects while (23) includes them. An analyst investigating the impact of pupil assignment and selection procedures based on grouping students by status would be stymied if he had to rely on information collected at the school level.

In contrast, unique estimates of interactive contextual effects can be obtained from aggregate data when

the effects arise from individual level variables included in the individual level relation. At the aggregate level, equation (20) becomes:

$$\bar{Y} = b_3\bar{X}_1 + b_4\bar{X}_1\bar{X}_2 + b_2\bar{X}_2 + e \quad (25)$$

while equation (21) is nonlinear:

$$\bar{Y} = b_1\bar{X}_1 + b_3\bar{X}_2 + b_4\bar{X}_2^2 + e. \quad (26)$$

To summarize, when contextual effects originate from individual level variables included in the individual level model, aggregate data cannot always yield unique estimates of those effects. Aggregate data are satisfactory if the effects are known to be interactive. However, the investigator's most likely task is to determine whether contextual effects are interactive, direct, or nonexistent. For that task, aggregate data are inadequate.

Aggregate data suffice for the analysis of contextual variables defined independently of variables included in the individual level relation. Consider, for instance, the effects of variables like the quality of teachers or school facilities, denoted by $\bar{Z}$. If $\bar{Z}$ has direct effects on student achievement, apart from the effects of race and socioeconomic status, then:

$$Y = b_1X_1 + b_2X_2 + b_3\bar{Z} + e \quad (27)$$

is an appropriate individual level model. At the aggregate level, the model still has three knowns and three unknowns:

$$\bar{Y} = b_1\bar{X}_1 + b_2\bar{X}_2 + b_3\bar{Z} + e. \quad (28)$$

Even if the effects of $\bar{Z}$ are interactive, an investigator can rely on aggregate data to secure unique estimates. Suppose that high quality school facilities mitigate the effects of race on achievement. Interracial achievement differentials, measured by b_1 , will therefore be less in schools with good facilities. Such a model is represented by:

$$Y = b_1X_1 + b_2X_2 + e,$$

where $b_1 = (b_3 + b_4\bar{Z})$, so that the full equation is:

$$Y = b_3X_1 + b_4X_1\bar{Z} + b_2X_2 + e. \quad (29)$$

The corresponding aggregate equation has identical form:

$$\bar{Y} = b_3\bar{X}_1 + b_4\bar{X}_1\bar{Z} + b_2\bar{X}_2 + e. \quad (30)$$

Thus, aggregate data alone are sufficient to estimate all three coefficients uniquely.

Aggregate data can also generate unique estimates if school facilities interact with socioeconomic status. Attempts to upgrade school facilities imply that the link between status and achievement can be diminished. Status differentials should therefore be reduced in high quality schools. This expectation is represented by:

$$Y = b_1X_1 + b_2X_2 + e,$$

where $b_2 = (b_3 + b_4\bar{Z})$, yielding a full model:

$$Y = b_1X_1 + b_3X_2 + b_4X_2\bar{Z} + e. \quad (31)$$

The aggregate version of this model has the same number of knowns as unknowns:

$$\bar{Y} = b_1\bar{X}_1 + b_3\bar{X}_2 + b_4\bar{X}_2\bar{Z} + e.$$

Thus, aggregate level data are sufficient to analyze direct as well as interactive contextual effects defined independently of individual level variables.

In sharp contrast, when contextual effects depend on values of the individual level dependent variable, unique estimation from aggregate data is impossible, no matter whether the effects are direct or interactive. Determining how the achievement of peers affects the achievement of individual students is crucial for the evaluation of policies based on tracking or ability grouping. Relying on aggregate data for such an evaluation, however, is likely to produce misleading conclusions.

First, suppose that peer achievement, represented by $\bar{Y}$, has direct effects on the achievement of individual students, apart from the effects of race and status:

$$Y = b_1 X_1 + b_2 X_2 + b_3 \bar{Y} + e. \quad (32)$$

In the absence of data on students, the corresponding school level model becomes:

$$\begin{aligned} \bar{Y} &= b_1 \bar{X}_1 + b_2 \bar{X}_2 + b_3 \bar{Y} + e \\ &= \frac{b_1}{(1-b_3)} \bar{X}_1 + \frac{b_2}{(1-b_3)} \bar{X}_2 + e. \end{aligned} \quad (33)$$

Equation (33) has more unknowns than knowns. Unless some a priori value is assigned to b_2 , the coefficient for status, unique estimation of b_3 , the coefficient for peer ability, is impossible. Thus, using data on schools renders estimates of the effects of peer ability on

individual achievement moot. Moreover, equation (33) is indistinguishable from both equation (22), which represents the direct effects of peer status, and equation (24), which represents the absence of all contextual effects. Thus, using aggregate data impedes evaluation of the alternative explanations of achievement that each of these equations depict.

If peer ability has interactive effects on achievement, aggregate data makes analysis completely intractable. Suppose, for instance, that the difference between the achievement scores of black and white students diminishes as the proportion of high achievers in their school increases. Controlling for race and socioeconomic status, such a relation is represented by:

$$Y = b_1X_1 + b_2X_2 + e$$

where $b_1 = (b_3 + b_4\bar{Y})$. By substitution, the full model becomes:

$$Y = b_3X_1 + b_4X_1\bar{Y} + b_2X_2 + e. \quad (34)$$

In the absence of data on students, the equivalent regression is:

$$\begin{aligned} \bar{Y} &= b_3\bar{X}_1 + b_4\bar{X}_1\bar{Y} + b_2\bar{X}_2 + e \\ &= \frac{b_3\bar{X}_1}{(1-b_4\bar{X}_1)} + \frac{b_2\bar{X}_2}{(1-b_4\bar{X}_1)} + e \end{aligned} \quad (35)$$

Equation (35) is not linear in its parameters. Therefore, conventional regression techniques yield no solution.

In addition, without additional assumptions about the value of either b_2 or b_3 , no estimation procedure at all could produce unique estimates of b_4 , the parameter representing the effects of peer ability on individual achievement. In summary, relying on aggregate data makes unique estimation impossible if peer ability has direct effects. If it has interactive effects, conventional procedures make all estimation impossible.

In summary, aggregate data yield unique estimates of the effects of contextual variables only under certain conditions. The cost and difficulty of obtaining data on students make it imperative to determine which types of contextual effects are present. If contextual properties like school facilities and teacher quality, defined independently of individual level variables, are the only significant contextual factors, then school level data may suffice to study the behavior of students. School level data are also adequate if contextual variables like peer status, originating from values of an independent variable included in the individual level relation, have interactive effects on achievement. Under other conditions, data on students will be necessary, particularly if educational research is to be relevant for educational

policy. Unique estimates of the impact of contextual variables are crucial for policy research. Only contextual properties are potentially subject to manipulation; properties of individuals are not.

III.3 Ecological Regression and the Effects of Grouping

This chapter has emphasized the importance of ecological regression for recovering estimates of individual level relations in the absence of individual level data. If the ecological model is properly specified, regardless of whether correct specification requires multiple independent variables or a nonlinear model, ecological regression will yield unbiased estimates of individual level relations. Some nonlinear models will also produce unique estimates of individual and contextual effects.

However, ecological data are data grouped by proximity. In any given instance, proximity grouping may approximate grouping by the independent variable, the dependent variable, or by both variables simultaneously. The last chapter showed that, unless data are grouped by the independent variable, regression estimates will be biased. How then can ecological regression in general yield unbiased regression estimates?

The next chapter discloses how this can occur. In brief, those grouping strategies which result in biased ecological regression estimates do so because the ecological models are themselves misspecified. These strategies, in effect, require inclusion of the grouping variables in the ecological regression. Failure to include these variables biases ecological regression estimates. Inclusion of these variables guarantees that ecological regression estimates will be unbiased, regardless of how the data were grouped. The following chapter then

reanalyses Coleman's (1966) data to show that the use of ecological regression to study student achievement appears to produce unbiased estimates. However, the presence of contextual effects precludes unique estimation. Thus, the empirical analysis in Ch. V uses the techniques discussed in this chapter as well as the next to argue that reliance on individual level data to study student achievement is essential if those studies are to have policy relevance.

IV. GROUPING STRATEGIES, ECOLOGICAL REGRESSION AND AGGREGATION BIAS^{*}

IV.1 Introduction

This chapter argues that aggregation bias can be analyzed as a special case of the general problem of properly specifying regression equations; the primary deficiency of previous work has been its failure to integrate the analysis of grouping procedures with the analysis of the specification problem. Two major, but thus far disparate, traditions can be discerned in the disaggregation literature. Each has been discussed in previous chapters. The first, deriving primarily from Leo Goodman's (1959) seminal work, concentrates on techniques for the use of regression rather than correlation as a means of bridging the gap between information about aggregates and statements about individuals. The second, deriving primarily from econometric studies, concentrates on the use of alternative grouping procedures to bridge this gap (Prais and Aitchison, 1954; Cramer, 1964; Feige and Watts, 1972). This chapter integrates the two traditions by focusing on the specification of the equations used at the individual and at the aggregate level of analysis. The first tradition ignores the question of grouping and

^{*}This chapter was co-authored with Allan J. Lichtman, Department of History, The American University.

assumes that equations are properly specified. The second tradition, while exploring the effect of grouping on aggregation bias, fail to recognize that such bias results from the influence of grouping on the specification of the aggregate level equation. The present study suggests that as long as the investigator uses regression rather than correlation measures, aggregation bias can be thoroughly understood by examining the specification of both the aggregate and individual level regression equations.

Aggregation bias results from changes in the proper specification of regression equations that are produced by grouping individuals into aggregate units. A regression equation is properly specified if the error term, representing unexplained variance, is unrelated to any of the independent variables included in the equation. This means that no variable has been excluded which is related both to the dependent variable and to one or more independent variables. The exclusion of such a variable biases the value of the regression coefficients for all independent variables with which it is related. The resulting specification bias is the difference between the biased and unbiased values of the coefficient being considered. Some forms of grouping alter the specification bias in the individual level regression model; others do not. Aggregation bias is generated by differences in the specification bias of individual and aggregate level regression estimates. Only if estimates at both levels of analysis are properly

specified or if they have equal specification bias will aggregation bias be zero. The only important exception arises when the individual behavior being considered is a function both of individual level variables and "contextual variables" which measure factors pertaining to the context or environment in which the behavior takes place. In this case, even with properly specified aggregate level regression models, it may not be possible to distinguish the unique effects of contextual and individual level variables.

When the assumption about proper specification at the individual level is incorrect, grouping can still create aggregation bias in the macrolevel regression model. Such aggregation bias may serve to either increase or decrease the specification bias in the individual level parameters. Thus if the researcher seeks estimates of individual level parameters which minimize specification bias, it may actually be beneficial to use aggregate rather than individual level data. When the individual level model includes contextual variables, however, properly specified aggregate level regression equations, while yielding unbiased estimates of individual level parameters, will not necessarily yield unique estimates. Only under certain conditions will it be possible to distinguish the influence of individual and contextual variables.

IV.2 Grouped Observations, Specification Error and Aggregation Bias - Correctly Specified Micromodels

The most general approach to showing how alternative grouping procedures and alternative specifications of micro and macrorelations affect the validity of grouped estimates is to use structural equations. The focus of this study is on the bias of regression coefficients, which are ratios of a covariance to a variance. Substituting structural equations into these terms and taking expectations discloses the differences (if any) between regression estimates from grouped and individual data. This section considers two cases: a correctly specified micro and macromodel, and a correctly specified micromodel in conjunction with an incorrectly specified macromodel.

Taking the first case, consider a perfectly specified microrelation of the form:

$$y_{ij} = \alpha + Bx_{ij} + u_{ij} \quad (1)$$

where $i = 1, \dots, t$ groups of as yet undefined structure, $j = 1, \dots, n_i$ observations within each group, and where u_{ij} has an expected value of zero, constant variance and is uncorrelated with x_{ij} . The notation for several variances and covariances is the same as that used in Chapter II.

The ordinary least squares estimate of B in equation (1) is given by:

$$b = \frac{s(x,y)}{s(x^2)} = \frac{\sum_{ij} (x_{ij} - \bar{x})(y_{ij} - \bar{y})}{s(x^2)} \quad (2)$$

Substituting equation (1) into equation (2) and taking expectations gives another, more useful expression for the expected value of the OLS estimate, b . As shown in Chapter II, substitution gives:

$$b = B + \frac{s(x,u)}{s(x^2)} \quad (3)$$

Therefore, the expected value of the microestimate is:

$$E(b) = B + E \left[\frac{s(x,u)}{s(x^2)} \right] \quad (4)$$

However, since the microrelation is assumed to be perfectly specified, $E(x,u) = 0$, and equation (4) reduces to:

$$E(b) = B$$

Now, suppose the data are grouped into t groups in some fashion. The macrorelation is given by:

$$\bar{y}_i = \alpha + B\bar{x}_i + \bar{u}_i \quad (5)$$

where B is the unknown microparameter, estimated by $E(\bar{b})$. This equation may or may not be perfectly specified. In fact, whether it is perfectly specified, so that the disturbance is uncorrelated with the independent variable, depends on how the data are grouped, as will be shown below. If the expression $\frac{s(\bar{x}, \bar{y})}{s(\bar{x}^2)}$ is used to calculate $\bar{b}$, then the logic given in (2) and (3) above means that:

$$\bar{b} = B + \frac{s(\bar{x}, \bar{\mu})}{s(\bar{x}^2)}.$$

Whenever equation (5) is perfectly specified, then $\bar{x}_i$ and $\bar{\mu}_i$ will be uncorrelated, so that $E[s(\bar{x}, \bar{\mu})] = 0$. Consequently, when both the macrorelation (5) and the microrelation (1) are perfectly specified,

$$E(b) = E(\bar{b}) = B. \quad (6)$$

In short, when microrelations and macrorelations are perfectly specified, there can be no aggregation bias. Estimates from grouped data will have the same expected value as estimates from ungrouped data, as long as the grouping procedure does not bias the macrorelation. As shown in Chapter II, two grouping procedures have this property: random grouping and grouping by x_{ij} .

The specification of the macrorelation, however, is affected by grouping the microdata according to values of y_{ij} , the dependent microvariable. If the highest values of y are placed in one group, the next highest in another, and so on until t groups are formed, x_{ij} and μ_{ij} will be correlated at the macrolevel even if they are uncorrelated at the microlevel. For example, if x_{ij} and y_{ij} are assumed to be positively related, and if y_{ij} and μ_{ij} are also assumed to be positively related, then placing high values of y in one group will also pull into that group many high values of both x and μ . Placing the next highest values of y in another group will pull in many of

the next highest values of x and u . Thus, at the grouped level, $\bar{x}_i$ and $\bar{u}_i$ will be correlated, so that $E(\bar{b}) = B + E[s(\bar{x}, \bar{u})]/s(\bar{x})$. The term $[E s(\bar{x}, \bar{u})]/s(\bar{x})$ is the specification bias in the macroestimate $\bar{b}$ that results from the confounding influence of $\bar{u}_i$. At the individual level x_{ij} and u_{ij} are uncorrelated, so that $E(b) = B$. The grouped estimate, $E(\bar{b})$, will be a biased estimate of the individual level parameter, $E(b)$. Since aggregation bias is the difference between the grouped and ungrouped estimates, its value is $E[s(\bar{x}, \bar{u})]/s(\bar{x})$. In this case, aggregation bias is equal to the specification bias in the macrorelation.

The source of the bias is not aggregation. Rather, the bias stems from misspecification of the macrorelation (5). The bias is properly viewed as specification bias, not as aggregation bias, contrary to the implications of many previous studies of this problem. However, it is possible to rewrite equation (5) by explicitly introducing a third variable, $\bar{z}_i$, which represents the confounding component of the micro error term u_{ij} created by grouping on y . When $\bar{z}_i$ is included in the macroequation, there is a new error term, $\bar{w}_i$, uncorrelated with both $\bar{x}_i$ and $\bar{z}_i$. Thus the new macrorelation is perfectly specified:

$$\bar{y}_i = \alpha + B_1 \bar{x}_i + B_2 \bar{z}_i + \bar{w}_i \quad (7)$$

At the microlevel, the independent effect of x_{ij} on y_{ij} is given by B ; B is estimated by b_{yx} . At the macrolevel, the effect is given by B_1 ; B_1 is estimated without bias by $\bar{b}_{yx.z}$. As a result, the expected value of the macroestimate $\bar{b}_{yx.z}$ will equal the expected value of the microestimate b_{yx} :

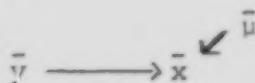
$$E(b_{yx}) = E(\bar{b}_{yx.z}) = B = B_1. \quad (8)$$

In other words, when specification bias is absent at both the micro and macrolevels, aggregation bias will also disappear. The aggregation bias which is said to result from grouping data by values of the dependent variable is an artifact due to misspecification in the original macrorelation (5).

Unfortunately there are no criteria for selecting an appropriate variable $\bar{z}_i$ for the respecification of macro-models that are biased by grouping on the dependent variable y . A more useful conceptualization of the problem rests upon the premise that the expected value of the aggregate level regression coefficient $E(\bar{b}_{yx})$ is biased because the grouping procedure has reversed the direction of the individual level relationship between x and y (Feige and Watts, 1972). If data are grouped into units according to values of y then y is necessarily the causal variable generating unit to unit variation in the values of other variables. The value of y for each unit is fixed

by the grouping process and the values of all other variables are a function of y . As Figure 1 illustrates, variations across units in the mean value of variable x will

Figure 1: Grouping by y .



result from the marginal effect of $\bar{y}$ on $\bar{x}$ ($\bar{b}_{xy}$). Thus for aggregate level data grouped on variable y it is appropriate to estimate $\bar{b}_{xy}$ rather than $\bar{b}_{yx}$. Moreover, just as $\bar{b}_{yx}$ equals b_{yx} when grouping on x , $\bar{b}_{xy}$ will equal b_{xy} since grouping on y will not affect the marginal influence of y on x by confounding y with the error term for x . Although at the individual level we are interested in recovering the true value of b_{yx} rather than b_{xy} , procedures have been developed for estimating b_{yx} from knowledge of the value for b_{xy} . As demonstrated in Chapter II, the regression coefficient b_{xy} can be used to estimate the individual level correlation coefficient r_{xy} . The value of b_{yx} can easily be computed from known values of b_{xy} and r_{xy} :

$$r_{xy}^2 = b_{xy} b_{yx}; b_{yx} = \frac{r_{xy}^2}{b_{xy}} \quad (9)$$

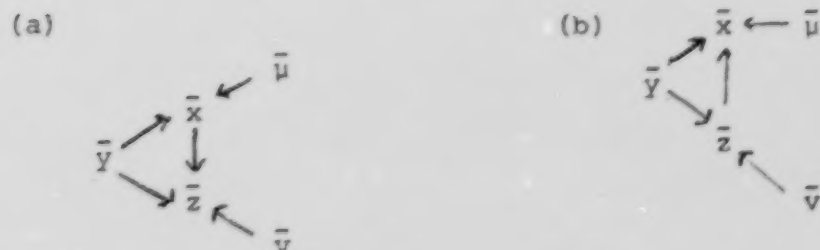
Theorists have not, however, addressed the problem of estimating individual level relationships from aggregate level information when data have been grouped on a

dependent variable y_i which is a multivariate function of several independent variables. The problem can be resolved by examining the structural relationships among a set of aggregate level variables when data are grouped on y_i . Assume without loss of generality that y_i is a perfectly specified function of x_i and z_i ;

$$y_i = a + b_1 x_i + B_2 z_i + u_i \quad (10)$$

If data are grouped on y the possible aggregate level relationships between y_i , x_i and z_i are portrayed in Figure 2.

Figure 2: Grouping by y : Multivariate Case.



Both $\bar{x}$ and $\bar{z}$ are functions of $\bar{y}$. Similarly, depending upon the direction of the individual level relationship between x_i and z_i , either $\bar{x}$ is also a function of $\bar{z}$, or $\bar{z}$ is also a function of $\bar{x}$. Given, however, that we are interested in recovering only the individual level influences of x_i and z_i on y_i , the direction of the causal relationship between x_i and z_i will not affect our argument. By regressing $\bar{x}$ on $\bar{y}$ and $\bar{z}$ we obtain $\bar{b}_{xy.z}$ which is a

perfectly specified aggregate level estimate the expected value of which is equivalent to the expected value of its corresponding individual level estimate: $E(\bar{b}_{xy.z}) = E(b_{xy.z})$. The individual level coefficient $b_{xy.z}$ can then be used to recover $b_{yx.z}$, the marginal influence of x_i on y_i controlling for z_i :

$$b_{yx.z} b_{xy.z} = r_{yx.z}^2$$

$$b_{yx.z} = \frac{r_{yx.z}^2}{b_{xy.z}} \quad (11)$$

Since $b_{xy.z}$ has already been estimated from $\bar{b}_{xy.z}$, knowledge of the value of the individual level partial correlation coefficient $r_{yx.z}^2$ would disclose the value of $b_{yx.z}$. Fortunately this partial correlation coefficient can be estimated from the formula for the multiple correlation $R_{y.xz}^2$:

$$R_{y.xz}^2 = r_{yz}^2 + r_{yx.z}^2 (1 - r_{yz}^2)$$

$$r_{yx.z}^2 = \frac{r_{y.xz}^2 - r_{yz}^2}{(1 - r_{yz}^2)} \quad (12)$$

Grouping on the independent variable does not affect the value of the individual level regression coefficient, so that $\bar{b}_{zy} = b_{zy}$. Using Goodman's procedure, b_{zy} can then be used to estimate r_{yz}^2 . Similarly $R_{y.xz}^2$ can be computed by the following formula:

$$R^2_{y.xz} = \frac{r^2_{yx} + r^2_{yz} - 2r_{yx}r_{yz}r_{xz}}{1 - r^2_{xz}} \quad (13)$$

Each term on the right hand side of (13) can be estimated directly or indirectly from the aggregate level regression coefficients. First, we have just shown that r^2_{yz} can be estimated from $\bar{b}_{zy}$. By the same logic, r^2_{yx} can similarly be estimated from $\bar{b}_{xy}$. Grouping on y , the independent variable, does not affect the value of the individual level regression coefficient, so that $\bar{b}_{xy} = b_{xy}$. Goodman's procedure discussed in Chapter II can then be used to derive r^2_{xy} .

The derivation of r_{xz} is more complex. Since the grouping variable y is independent of the error terms $\bar{u}$ and $\bar{v}$, $\bar{b}_{zx.y} = b_{zx.y}$ and $\bar{b}_{xz.y} = b_{xz.y}$. Further, because $\bar{b}_{zx.y} \times \bar{b}_{xz.y} = \bar{r}^2_{xz.y}$ and $b_{zx.y} \times b_{xz.y} = r^2_{xz.y}$ by definition, it follows from the preceding statement that $\bar{r}^2_{xz.y} = r^2_{xz.y}$. Again, by definition,

$$r^2_{xz.y} = \left[\frac{r_{xz} - r_{xy}r_{zy}}{\sqrt{1-r^2_{xy}} \sqrt{1-r^2_{zy}}} \right]^2.$$

The steps for computing $r^2_{xz.y}$, r_{xy} and r_{zy} have been outlined by the preceding statements, so the remaining unknown, r_{xz} , can be found. Since each term on the right hand side of equation (13) can be determined from the aggregate regression coefficients (together with knowledge

of the total sum of squares of each variable) $R^2_{y.xz}$ can also be determined.

Finally, having estimated the values of r^2_{yz} and $R^2_{y.xz}$, simple substitution of these values into the right hand side of equation (12) gives the value for $r^2_{yx.z}$. By equation (11), knowing the values of $b_{xy.z}$ and $r^2_{yx.z}$ is sufficient to estimate the coefficient of interest, $b_{yx.z}$. Thus, even when the properly specified aggregate relation includes multiple regressors, it is still possible to estimate the individual level influence of x_i on y_i as long as the total sum of squares of each variable is known.

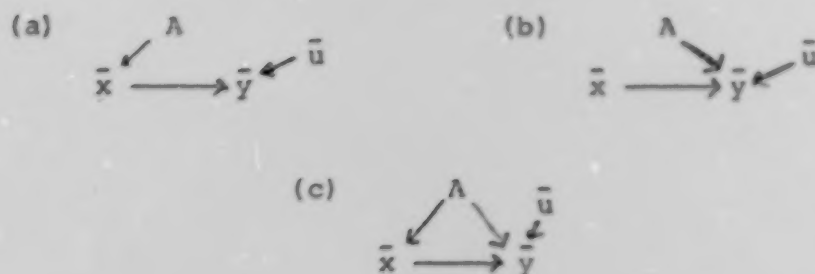
Thus the problems of disaggregation posed by grouping on y are far less intractable than previous work has suggested. If the investigator can properly specify the individual level regression model in either of the two ways suggested it is possible to estimate without aggregation bias the values of the individual level parameters in both the univariate and the multivariate case.

More common is the situation in which data are grouped by a variable related either to y_{ij} or to x_{ij} or to both. Consideration of this grouping procedure is especially important because it approximates grouping by geographic proximity, the form of aggregation most frequently encountered by policy analysts. When data are grouped by a variable, A , likely to be related directly to x_{ij} and y_{ij} , the macrorelation (5) may be misspecified by the omission

of the grouping criterion. If Λ is related to $\bar{x}$ alone, or if Λ is related to $\bar{y}$ alone, as shown in Figure 3a and 3b, then (5) is perfectly specified. If Λ is related to both $\bar{x}$ and $\bar{y}$, (5) is incorrectly specified. The proper model should include both x and Λ as independent variables. The most general form of the macrorelation in the case of proximity grouping will include Λ as well as x :

$$\begin{aligned}\bar{y}_i &= a + b_1 \bar{x}_i + B_2 \Lambda + \bar{u}_i \\ &= a + E(\bar{b}_{yx.\Lambda}) \bar{x}_i + E(\bar{b}_{y\Lambda.x}) \Lambda + \bar{u}_i\end{aligned}\quad (14)$$

Figure 3: Macrorelations: Grouping by Proximity.



Even if Λ is related only to $\bar{x}$ or only to $\bar{y}$, the macroestimate from (5), $B = E(\bar{b}_{yx})$ will equal the macroestimate of the influence of $\bar{x}$ on $\bar{y}$ generated by (14):

$$B_1 = E(\bar{b}_{yx.\Lambda}) = B = E(\bar{b}_{yx})$$

The corresponding micromodel also may not be properly represented by equation (1), as shown in Figure 4. Figure 4a indicates that the grouping variable Λ may be directly related to both x_{ij} and y_{ij} at the microlevel, in which case the correct micromodel is:

$$\begin{aligned} y_{ij} &= a + B_1 x_{ij} + B_2 \Lambda_j + u_{ij} \\ &= a + E(b_{yx.\Lambda}) x_{ij} + E(b_{y\Lambda.x}) \Lambda_j + u_{ij} \end{aligned} \quad (15)$$

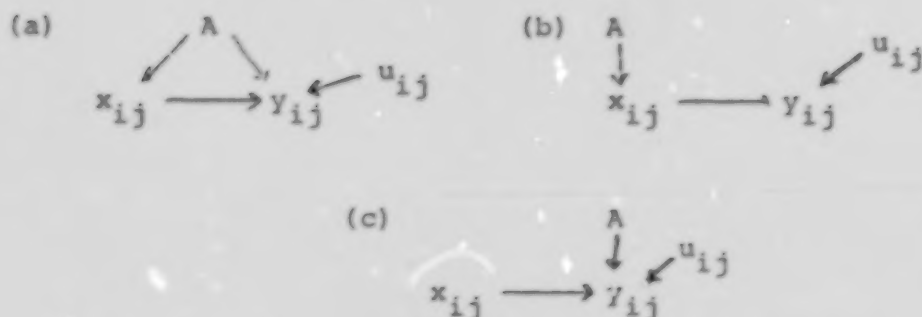
If Λ is directly related only to x_{ij} or if Λ is directly related only to y_{ij} , as shown in Figure 4b and 4c, respectively, Λ need not be included in the microrelation. Equation (1) will correctly represent the structural relation between x_{ij} and y_{ij} . However, under these circumstances, use of equation (15) rather than equation (1) will still yield unbiased estimates of the influence of x_{ij} on y_{ij} . In other words, B_1 from equation (15) will equal B from equation (1): $B_1 = E(b_{yx.\Lambda}) = B = E(b_{yx})$. Thus, equation (15) is the most general representation of the correctly specified micromodel, and it will be used in place of equation (1).

Use of the properly specified micromodel (15) and the properly specified macromodel (14) guarantees that aggregation bias will not occur. The independent effect of x on y at the microlevel is indicated by $E(b_{yx.\Lambda}) = B_1$. At the macrolevel, the same effect is given by $E(\bar{b}_{yx.\Lambda}) = B_1$. As a result, the expected values of the macro and microestimates

will be equivalent:

$$E(b_{yx.A}) = E(\bar{b}_{yx.A}) = B_1 \quad (16)$$

Figure 4: Possible Microrelations: Grouping by Proximity.



Thus, regardless of the grouping strategy, there will be no aggregation bias when both the micro and macroequations are properly specified.

Failure to properly specify the macrorelation by omitting the grouping variable A , as well as any other source of misspecification in the macromodel, results in aggregation bias equal to the specification bias in the macrorelation. More specifically, let equation (15) be a properly specified microrelation and let equation (14) be the proper macrorelation. Suppose, however, that the variable A is omitted from (14), so that equation (5) is used instead of (14). Use of equation (5) rather than (14) results in aggregation bias equal to the specification bias in (5), as shown by substituting (14) into the formula for estimating $\bar{b}_{yx}$ from equation (5) and taking expectations

(Draper and Smith, 1966:81-82):

$$\begin{aligned}
 E(\bar{b}_{yx}) &= \frac{E[S(\bar{x}, \bar{y})]}{s(\bar{x}^2)} \\
 &= \frac{E[\sum (\bar{x}_i - \bar{\bar{x}})(\bar{y}_i - \bar{\bar{y}})]}{s(\bar{x}^2)} \\
 &= \frac{E[\sum (\bar{x}_i - \bar{\bar{x}}) \{ (a + \bar{b}_{yx.A} \bar{x}_i + \bar{b}_{yA.x} \Lambda + \bar{u}_i) - (a + \bar{b}_{yx.A} \bar{\bar{x}} + \bar{b}_{yA.x} \bar{\Lambda} + \bar{\bar{u}}) \}]}{s(\bar{x}^2)} \\
 &= E[\sum (\bar{x}_i - \bar{\bar{x}}) \{ \bar{b}_{yx.A} (\bar{x}_i - \bar{\bar{x}}) + \bar{b}_{yA.x} (\Lambda - \bar{\Lambda}) + (\bar{u}_i - \bar{\bar{u}}) \} / s(\bar{x}^2)] \\
 &= E[\bar{b}_{yx.A} s(\bar{x}^2) + \bar{b}_{yA.x} s(\Lambda, \bar{x}) + s(\bar{x}, \bar{u})] / s(\bar{x}^2) \\
 &= E(\bar{b}_{yx.A}) + E(\bar{b}_{yA.x}) \bar{b}_{Ax} + \frac{E[s(\bar{x}, \bar{u})]}{s(\bar{x}^2)} \quad (17)
 \end{aligned}$$

The expected value of the last term in (17), however, is zero since $\bar{x}_i$ and $\bar{u}_i$ are uncorrelated in equation (14). The expected value of the influence of $\bar{x}_i$ on $\bar{y}_i$ is therefore:

$$E(\bar{b}_{yx}) = E(\bar{b}_{yx.A}) + E(\bar{b}_{yA.x}) \bar{b}_{Ax} \quad (18)$$

The specification bias in the macrorelation (5) is the difference between the biased and unbiased expectations:

$$E(\bar{b}_{yx}) - E(\bar{b}_{yx.A}) = E(\bar{b}_{yA.x}) \bar{b}_{Ax} \quad (19)$$

Because the microrelation (15) is properly specified, its specification bias is zero: $E(b_{yx.A}) - E(b_{yx.A}) = 0$. Since aggregation bias is the difference between the expected value of the macro and microestimates of the influence of x on y generated by the equations actually used in the analysis, it can be expressed by:

$$\begin{aligned} E(\bar{b}_{yx}) - E(b_{yx.A}) &= E(\bar{b}_{yx.A}) + E(\bar{b}_{yA.x})\bar{b}_{Ax} - E(b_{yx.A}) \\ &= E(\bar{b}_{yA.x})\bar{b}_{Ax} \end{aligned} \quad (20)$$

Therefore, because the right hand side of (19) and (20) are equal, the aggregation bias that results from misspecifying macrorelations when microrelations are properly specified is equal to the specification bias in the misspecified macroequation.

In sum, when the individual level equation is properly specified, there will be no aggregation bias in parameter estimates if the group level equation is also properly specified. Aggregation bias occurs only when the macro-model is improperly specified, and is equal to the specification error in the macroestimate. Specification error in macroestimates can be created by grouping variables which, if omitted, will confound independent variables with the error term of the improperly specified microrelation. Through respecification of the aggregate model, however, it is still possible to eliminate this confounding influence and obtain unbiased estimates of individual

level parameters. When dealing with properly specified microrelations, as long as there is no specification error in aggregate level equations, there will be no aggregation error, regardless of how the data are grouped.

Properly specifying relations in order to avoid aggregation bias may be quite difficult in practice. Aside from the fact that if correctly specified models were known, additional research would be otiose, there are other less obvious difficulties. Some of these have already been mentioned for the case of grouping by y , where the impracticality of defining and measuring the systematic component of the aggregate error term seems clear. When data have been grouped by a variable related to both x and y , there are two additional problems which are particularly troublesome because proximity grouping is so common and is likely to approximate data grouped in this manner. The first problem is that of "controlling for A " or bringing A explicitly into the regression equation, when each category of A represents a unique geographical unit. The second problem is the potentiality of non-linear relations between a geographic grouping variable and the other variables included in the macroregression.

When each category of A represents a unique geographical unit, its inclusion in a multiple regression equation is awkward at best and impossible to interpret at worst. To define A , it is first necessary to understand what mechanisms make individuals come together in

groups. For example, suppose an analyst wishes to study the relationship between race and the propensity to commit violent crimes, but he has data only at the county level. He cannot include each county as a dummy independent variable in the macrorelation, because he would then have more independent variables (3130 counties in the United States and one other independent variable) than observations ($N = 3130$ counties). He must define what property, if any, of counties makes high proportions of nonwhites appear in the same counties with high rates of violent crime. Several properties suggest themselves: urbanism, high density, inadequate employment opportunities. Undoubtedly, the attraction that certain counties have for both nonwhites and criminals is a combination of these as well as other properties. In short, while "proper names" do not suffice, specifying the process that places individuals into groups may also be extremely difficult.

Especially vexing is the situation in which aggregate level variables are a nonlinear function of the grouping criterion. Following standard practice, this study of the aggregation problem assumes that $\bar{x}$ and $\bar{y}$ are linear functions of the grouping criterion A . The grouping of data according to nonlinear functions invalidates the argument that aggregation bias vanishes when both micro and macrorelations are properly specified. Aggregation bias will still be a problem when A is related to both $\bar{x}$

and $\bar{y}$ and one or both of these relations is nonlinear and when A is related in nonlinear fashion to either $\bar{x}$ or $\bar{y}$. Assume, for example, that an individual's welfare payments (y_i) are a perfectly specified function of his income (x_i). Assume further that data are grouped according to a nonlinear function of x such that the wealthiest individuals are included in the same group with the poorest individuals, the next wealthiest with the next poorest, etc. This hypothetical case may approximate many real world situations especially in less developed nations. At the aggregate level $\bar{y}$ is still a perfectly specified function of $\bar{x}$. Unfortunately, the nonlinear grouping has fundamentally altered the nature of the relationships between x and y so that $\bar{b}_{yx} \neq b_{yx}$. If the poorest individuals receive the largest welfare payments, b_{yx} will be positive, but $\bar{b}_{yx}$ may actually be negative, since the counties with the greatest proportion of poor people may have the highest mean income due to the presence of a few very wealthy individuals.

Detailed consideration of the problem of nonlinear grouping is beyond the scope of this article. We can, however, briefly mention several partial solutions designed to approximate the underlying individual level relationships that the investigator seeks to estimate. First, the investigator can use measures (for example the median or the tri-mean) that are more resistant than the mean to the presence of extreme values. Second, if the investigator

has information about within group distributions he can drop extreme values. Third, he can approximate linearity by analyzing subgroups which are approximately linear with respect to the grouping criterion. For example, in the case of skewed income distributions the investigator can regress mean welfare payments not on mean income but on the percentages of individuals in a county with incomes between 0 and \$5,000; between \$5,001 and \$10,000 and so on.

IV.3 Grouped Observations, Specification Error and Aggregation Bias: Incorrectly Specified Micromodels

In this section, misspecified individual level models are considered in conjunction with aggregate models that are both misspecified and correctly specified. Consideration of these models permits identification of the conditions under which aggregation bias amplifies specification bias as well as the conditions under which using aggregate level models can reduce specification bias in an original micromodel. When aggregation bias amplifies specification bias, an aggregation loss has occurred; when it reduces specification bias, an aggregation gain has occurred. Aggregation gain can result from a macromodel that is better specified than the corresponding micromodel. Under certain circumstances, aggregation gain may also result from the grouping process itself, even when the aggregate model is not better specified.

First consider y to be a function of two independent variables, one of which is excluded from the regression model. In turn, grouping by the included independent variable, the excluded independent variable, and a grouping criterion related to both independent variables will be examined. Second, y will be considered as a function of three independent variables, two of which are excluded, and the effects of grouping on one of the excluded independent variables will be considered.

Once again, these arguments will be elaborated through the use of structural equations to identify the specification bias in the individual and aggregate level relations. Since aggregation bias is the difference in the expected value of the estimated influence of x on y given by the aggregate and individual relations (*i.e.*, $E(\bar{b}_{yx}) - E(b_{yx})$), identification of the specification bias in both micro and macrorelations also permits identification of aggregation bias. Whether aggregation gain or loss occurs can then be determined. If specification bias in an individual relation is negative, so that the computed value of b_{yx} underestimates the true parameter, there will be an aggregation gain when the aggregation bias is positive. Similarly, if specification bias at the microlevel is positive, there will be an aggregation gain when the aggregation bias is negative. Of course, when there is no specification bias at the microlevel, aggregation bias will always produce aggregation loss.

Analysis of the three variable model relies on a properly specified multivariate microrelation of the form:

$$y_i = a + B_{yx.z} x_i + B_{yz.x} z_i + \mu_i \quad (21)$$

where $E(\mu_i) = 0$ and $E[\text{Cov}(x_i, \mu_i)] = E[\text{Cov}(z_i, \mu_i)] = 0$.

Let equation (21) be misspecified by the omission of z_i :

$$y_i = a + B_{yx} x_i + \mu'_i \quad (22)$$

Use of equation (22) rather than (21) results in specification bias, shown by substituting equation (21) into the formula for estimating B_{yx} and taking expectations:

$$\begin{aligned} E(b_{yx}) &= \frac{E[s(x,y)]}{s(x^2)} \\ &= \frac{E[\sum (x_i - \bar{x})(y_i - \bar{y})]}{s(x^2)} \\ &= \frac{E\{\sum (x_i - \bar{x}) [(a + B_{yx.z} x_i + B_{yz.x} z_i + \mu_i) - (a + B_{yx.z} \bar{x} + B_{yz.x} \bar{z} + \bar{\mu}_i)]\}}{s(x^2)} \\ &= \frac{E\{\sum (x_i - \bar{x}) [B_{yx.z} (x_i - \bar{x}) + B_{yz.x} (z_i - \bar{z}) + (\mu_i - \bar{\mu}_i)]\}}{s(x^2)} \\ &= \frac{E[B_{yx.z} s(x^2) + B_{yz.x} s(x,z) + s(x,\mu)]}{s(x^2)} \end{aligned}$$

$$\begin{aligned}
&= B_{yx.z} + B_{yz.x} \frac{E[s(x,z)]}{s(x^2)} + \frac{E[s(x,u)]}{s(x^2)} \\
&= B_{yx.z} + B_{yz.x} b_{zx} + \frac{E[s(x,u)]}{s(x^2)} \quad (23)
\end{aligned}$$

The expected value of the last term in (23), however, is zero, since x_i and u_i are uncorrelated in equation (21). The expected value of the estimated influence of x_i on y_i is therefore:

$$E(b_{yx}) = B_{yx.z} + B_{yz.x} b_{zx} \quad (24)$$

Specification bias in the microequation (22) is simply the difference between the expected value of b_{yx} and the true value, given by $B_{yx,z}$:

$$E(b_{yx}) - B_{yx,z} = B_{yz.x} b_{zx} \quad (25)$$

Assume now that the aggregate data have been grouped by x , the independent variable included in the misspecified microrelation (22). It was seen in the previous section that the properly specified macrorelation which corresponds to the microrelation (21) when data are grouped by an independent variable is:

$$\bar{y}_i = a + \bar{B}_{yx.z} \bar{x}_i + \bar{B}_{yz.x} \bar{z}_i + \bar{u}_i \quad (26)$$

Assume additionally that the macrorelation is misspecified in the same way that (22) is misspecified; that is, $\bar{z}_i$ is omitted:

$$\bar{y}_i = a + \bar{B}_{yx} \bar{x}_i + \bar{u}_i \quad (27)$$

By the same logic used in equation set (23), the specification bias at the aggregate level, when data are grouped by x , is:

$$E(\bar{b}_{yx}) - \bar{B}_{yx.z} = \bar{B}_{yz.x} \bar{b}_{zx} \quad (28)$$

To determine whether aggregation gain can occur in this instance it is necessary to consider the equation for aggregation bias:

$$E(\bar{b}_{yx}) - E(b_{yx}) = (\bar{B}_{yx.z} + \bar{B}_{yz.x} \bar{b}_{zx}) - (B_{yx.z} + B_{yz.x} b_{zx}) \quad (29)$$

As demonstrated in the previous section, however, there is no aggregation bias when both micro and macro-estimates are properly specified. Thus $\bar{B}_{yx.z} = B_{yx.z}$ and $\bar{B}_{yz.x} = B_{yz.x}$. By cancelling and factoring equivalent terms, equation (29) can be reduced to the following form:

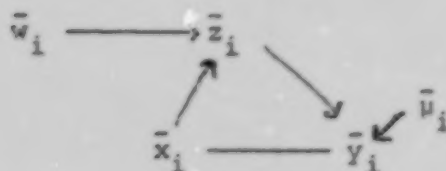
$$E(\bar{b}_{yx}) - E(b_{yx}) = B_{yz.x} (\bar{b}_{zx} - b_{zx}) \quad (30)$$

Equation (30) reveals that aggregation gain will occur only when aggregation bias, expressed by $B_{yz.x}(\bar{b}_{zx} - b_{zx})$, is in the opposite direction of specification bias in the microestimate, expressed by $B_{yz.x} b_{zx}$.

To determine the sign relationship between $B_{yz.x}(\bar{b}_{zx} - b_{zx})$ and $B_{yz.x} b_{zx}$, two special cases must be considered. First, assume that the underlying causal model represented by

(26) is that portrayed in Figure 5, where $\bar{x}_i$ influences $\bar{z}_i$, but $\bar{z}_i$ does not affect $\bar{x}_i$. In this case, $\bar{z}_i$ is an endogenous variable and $\bar{x}_i$ is exogenous. Thus, $\bar{x}_i$ is an independent variable with respect to $\bar{z}_i$, and grouping the data by x will not confound values of $\bar{x}_i$ with values of the disturbance term associated with z_i . The values of $\bar{b}_{zx}$ and b_{zx} will be equal and aggregation bias ($B_{yz.x} - (\bar{b}_{zx} - b_{zx})$) will be zero. Without aggregation bias, of

Figure 5: Properly Specified Macromodel, Data Grouped by x : Case I.

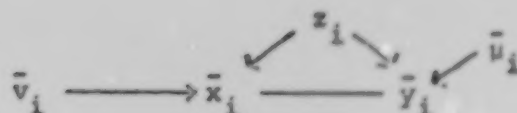


course, there is no possibility of aggregation gain (or loss), regardless of whether the specification bias represented by (28) is positive or negative.

The second special case also assumes that both individual and aggregate relations are misspecified by the omission of z and that data are grouped by x . In contrast to the first case, it assumes that $\bar{z}_i$ influences $\bar{x}_i$, but $\bar{x}_i$ does not influence $\bar{z}_i$, as illustrated by Figure 6. $\bar{z}_i$ is now an exogenous variable, while $\bar{x}_i$ is endogenous; $\bar{v}_i$ is its disturbance term. Since $\bar{x}_i$ is now dependent on $\bar{z}_i$, grouping data by x will confound the relation between $\bar{z}_i$ and $\bar{v}_i$, the error term associated

with $\bar{x}_i$. As a consequence, the corresponding terms on the right side of the equal sign in equation (33), $\bar{b}_{zx}$ and b_{zx} , will no longer be equal. Grouping by x in this case is equivalent to grouping by a dependent variable,

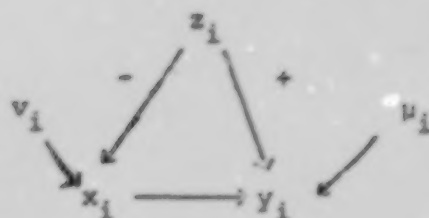
Figure 6: Properly Specified Macromodel, Data Grouped by x : Case II.



because x is dependent on z . At the aggregate level, values of $\bar{v}_i$ will not be independent of values of $\bar{z}_i$. Hence the microestimate $\bar{b}_{zx}$ will not equal the microestimate b_{zx} , since $\bar{v}_i$ and $\bar{z}_i$ are confounded in the aggregate but not in the individual model.

To assess whether or not there can be aggregation gain or loss in this case, consider first the situation in which specification bias in the individual level relation is negative, as shown by the signs in Figure 7. Negative specification bias means that $B_{yz.x} b_{zx} < 0$. For an aggregation gain, aggregation bias must therefore be positive to remove the effects of underestimation due to negative misspecification. In other words, $B_{yz.x}(\bar{b}_{zx} - b_{zx})$ must be greater than zero. Given the model in Figure 7, where b_{zx} is assumed to be negative and $B_{yz.x}$ is assumed to be positive, there are two ways

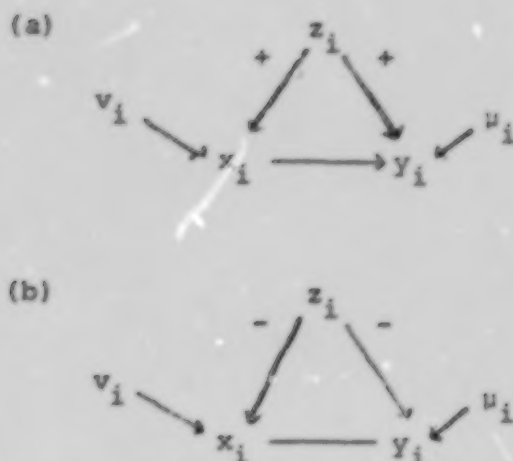
Figure 7: Negative Specification Bias--Micromodel.



that this condition can be fulfilled. The term could be greater than zero if $\bar{b}_{zx} > 0$ or if $\bar{b}_{zx}$ is both negative and less than b_{zx} . Neither of these results, however, is possible. Grouping on x_i brings together high values of v_i with low values of z_i , and low values of v_i with high values of z_i . Such grouping can only increase the already negative relationship between x_i and z_i ; $\bar{b}_{zx}$ can neither have a positive sign nor be smaller in magnitude than b_{zx} . Consequently, aggregation gain is impossible; only aggregation loss can occur.

Consider now the situation in which specification bias in the individual level relation is positive, as shown in Figure 8a and 8b. Positive specification bias means that $B_{yz.x} b_{zx} > 0$. Using data grouped by x will counteract this bias only if aggregation bias is negative: $B_{yz.x} (\bar{b}_{zx} - b_{zx}) < 0$. If $B_{yz.x}$ and b_{zx} are both positive, as they are in Figure 8a, aggregation bias will be negative under two circumstances. It will be negative if $\bar{b}_{zx} < 0$, or if $\bar{b}_{zx}$ is both positive and less than b_{zx} . Again, neither situation can occur since grouping data on

Figure 8: Positive Specification Bias--Micromodel.



x will increase the already positive relation between x and z by bringing together high values of v_i with high values of z_i , and low values of v_i with low values of z_i . Once again, only aggregation loss is possible.

If $B_{yz.x}$ and b_{zx} are both negative, as shown in Figure 8b, the same conclusion holds. The term $B_{yz.x}(\bar{b}_{zx} - b_{zx})$ could be negative if $\bar{b}_{zx} > 0$ and $\bar{b}_{zx} < b_{zx}$. For now familiar reasons, both these situations are impossible; grouping on x_i will only enhance the already negative relationship between x_i and z_i .

To summarize, when both the aggregate and the individual level relations are similarly misspecified, and when data are grouped on an independent variable included in both of the misspecified models, aggregation gain is impossible. If the properly specified group level model is that shown in Figure 5, where the data are grouped on

the exogenous independent variable, there will be neither an aggregation gain nor a loss. However, when data are grouped on an endogenous independent variable, illustrated by Figure 6, there can only be an aggregation loss; the process of aggregation will always amplify specification error.

It can further be shown that grouping data by z , the variable omitted from the misspecified macro and micromodels also produces aggregation loss but not aggregation gain. Since z is an omitted but confounding variable, grouping by z is equivalent to grouping the data by a variable directly related to both x and y , a case introduced in the previous section. To show that grouping by z makes aggregation gain impossible, it is first necessary to identify the specification bias in the micro and macroequations. The properly specified individual level relation is given by (21); the misspecified one is given by (22). Thus, its specification bias is given by (25). Similarly, the properly specified group level relation is represented by (26), and the misspecified relation is represented by (27). Thus, aggregation bias and specification error are denoted by the same terms as the case of grouping by x . (See equations (30) and (28) respectively.)

The possibility of aggregation gain when data are grouped by z and the individual and group relations are

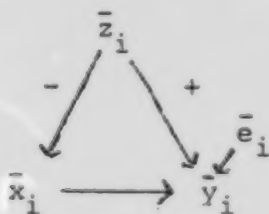
misspecified by the omission of z can be explored by first examining the case where specification bias in the individual relation is negative; i.e., when

$B_{yz.x} b_{zx} < 0$. (See Figure 9.) To obtain aggregate estimates which are closer to the true parameter than those in the misspecified individual model, aggregation bias must be positive:

$$B_{yz.x} (\bar{b}_{zx} - b_{zx}) > 0. \quad (31)$$

It has already been shown that this term must always be negative, given the posited conditions. Thus, since $B_{yz.x}$ is positive, $(\bar{b}_{zx} - b_{zx})$ must be negative. Given that b_{zx} is assumed to be negative, grouping on z can only increase the already negative relation between z and x . Consequently, $(\bar{b}_{zx} - b_{zx}) < 0$. Aggregation gain is thus impossible; only aggregation loss can occur.

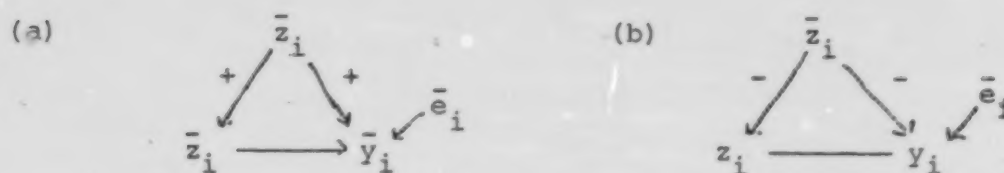
Figure 9: Perfectly Specified Macromodel: Grouping by z , Negative Specification bias.



A similar argument prevails when specification bias is positive. When specification bias in the individual relation is positive, aggregation gain requires

aggregation bias to be negative. Figure 10 shows the two conditions under which specification bias might be positive. In (a), the relations between $\bar{z}_i$ and $\bar{y}_i$ and between $\bar{z}_i$ and $\bar{x}_i$ are both positive; in (b), the relations between $\bar{z}_i$ and $\bar{y}_i$ and between $\bar{z}_i$ and $\bar{x}_i$ are both negative. Discussion of each individual model seems unnecessary, since the outcomes have a similar pattern. For aggregation gain to occur when specification bias is positive, equation (31) must be negative. It has already been seen however, that (31) cannot be negative under the posited conditions of Figure 10, because $B_{yz.x} > 0$ and $\bar{b}_{zx} > b_{zx}$.

Figure 10: Perfectly Specified Macromodel: Grouping by z , positive Specification Bias.



Once again, aggregation gain cannot occur.

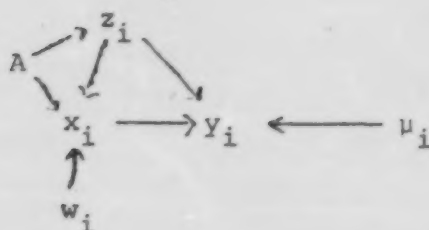
The next case to be considered involves grouping on neither z_i , a variable omitted from the microequation, nor x_i , a variable included in the microequation, but on a grouping criterion (A) that is related to both these independent variables. This situation, portrayed in Figure 11, is especially important both because proximity grouping generally involves aggregation on more than a single variable and because both aggregation loss

and aggregation gain can occur when data are grouped by A.

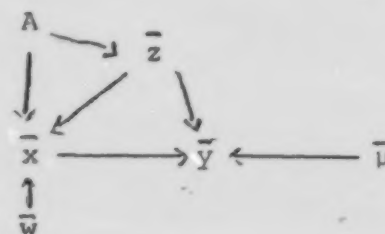
To determine whether aggregation gain or loss stems from grouping on A it is necessary to determine the expressions for specification bias in the microestimate

Figure 11: Grouping by A.

(a) Micromodel



(b) Macromodel



(b_{yx}) and for aggregation bias in the macroestimate $(\bar{b}_{yx})$. Given, however, that A has no direct effect on y it need not be included in the structural equation for y or in the expressions for specification bias in the micro-model and aggregation bias in the macromodel. Thus these expressions are equivalent to those in the cases of grouping by x or z. Specification bias in the microestimate equals $B_{yz.x} b_{zx}$ and aggregation bias in the macroestimate equals $B_{yz.x} (\bar{b}_{zx} - b_{zx})$. If b_{zx} and $(\bar{b}_{zx} - b_{zx})$ have the same signs specification bias and aggregation bias will also have the same signs and aggregation loss will occur; if they have opposite signs aggregation gain will occur. The expected values of b_{zx} and $\bar{b}_{zx}$ are expressed by the following equations:

$$E(b_{zx}) = B_{zx.A} + B_{zA.x}b_{Ax} \quad (32)$$

$$E(\bar{b}_{zx}) = B_{zx.A} + B_{zA.x}\bar{b}_{Ax} \quad (33)$$

Therefore:

$$E(\bar{b}_{zx} - b_{zx}) = B_{zA.x}(\bar{b}_{Ax} - b_{Ax}) \quad (34)$$

Equation (34) now has a familiar form. It is equivalent to the equation for aggregation bias in the macroestimate when grouping by an excluded variable related both to an independent and the dependent variables (see equation 31). The only difference is that A rather than z is now the excluded variable and z rather than y is the dependent variable. It has already been demonstrated, however, that the equation for aggregation bias in the case of grouping on an excluded confounding variable always has the same sign as the equation for specification bias in the microestimate. Since $B_{zA.x}b_{Ay}$ represents specification bias in the microestimate b_{zx} , $B_{zA.x}b_{Ax}$ will always have the same sign as $E(\bar{b}_{zx} - b_{zx})$.

Given that aggregation gain or loss in the macroestimate $\bar{b}_{yx}$ depends solely on the sign relationship between $E(\bar{b}_{zx} - b_{zx})$ and $E(b_{zx})$ and that $E(\bar{b}_{zx} - b_{zx})$ always has the same sign as $B_{zA.x}b_{Ax}$, we need focus our attention solely on the sign relationship between $E(b_{zx})$ and $B_{zA.x}b_{Ax}$. If $B_{zA.x}b_{Ax}$ and $E(b_{zx})$ have the same signs then $E(\bar{b}_{zx} - b_{zx})$ and $E(b_{zx})$ will likewise have the same

signs and only aggregation loss can occur. According to equation (32), the condition for aggregation loss (i.e., $E(b_{zx})$ has the same sign as $B_{zA.x}b_{Ax}$ and thus as $E(\bar{b}_{zx} - b_{zx})$) will be fulfilled when $B_{zx.A}$ and $B_{zA.x}b_{Ax}$ have the same sign and when they have opposite signs but $B_{zA.x}b_{Ax}$ has a larger magnitude than $B_{zx.A}$. The condition for aggregation gain (i.e., $E(b_{zx})$ has the opposite sign from $B_{zA.x}b_{Ax}$ and thus from $E(\bar{b}_{zx} - b_{zx})$) will be fulfilled when $B_{zx.A}$ and $B_{zA.x}b_{Ax}$ have opposite signs and $B_{zA.x}$ has a larger magnitude than $B_{zx.A}b_{Ax}$.

Thus grouping on a variable related to an included and an excluded independent variable will generate aggregation loss if such grouping amplifies the relationship between the included and excluded variables and aggregation gain if it diminishes the relationship between these two variables. This finding has important implications for investigators with the option to group data according to preset criteria. To reduce specification bias in parameter estimates, data should be grouped by variables that diminish the association between included and confounding independent variables. If such grouping variables can be used, aggregate data may yield more accurate estimates of direct individual level relationships than individual level data.

Additionally, in the case of a micromodel which excludes two or more independent variables it is possible to achieve an aggregation gain when grouping on one of

the excluded variables. Suppose, for instance, that the fully specified microrelation is:

$$y_i = a + B_{yx.zv}x_i + B_{yz.xv}z_i + B_{yv.xz}v_i + \mu_i \quad (35)$$

where $B_{yz.xv}$ and $B_{yv.xz}$ are assumed to be positive.

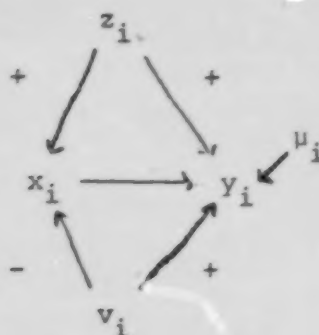
Further, v_i is assumed to have a negative influence on x_i and the influence of v_i on x_i and y_i is assumed to be less than that of z_i . (See Figure 12.) Suppose in addition that an analyst omits v_i and z_i , using the equation:

$$y_i = a + B_{yx}x_i + \mu_i \quad (36)$$

in place of equation (35). The specification bias that results from using (36) rather than (35) to estimate the independent influence of x_i on y_i is:

$$E(b_{yx}) - B_{yx.zv} = B_{yz.xv}b_{zx} + B_{yv.xz}b_{vx} \quad (37)$$

Figure 12: Properly Specified Micromodel.



Since the confounding influence of z_i is presumed to be greater than the confounding influence of v_i , this specification bias will always be positive. In lieu of microdata, the analyst uses data grouped by v together with a misspecified macrorelation:

$$\bar{y}_i = a + \bar{B}_{yx} \bar{x}_i + \bar{\mu}_i \quad (38)$$

The proper macrorelation must include both $\bar{z}_i$ and $\bar{v}_i$:

$$\bar{y}_i = a + \bar{B}_{yx.zv} \bar{x}_i + \bar{B}_{yz.xv} \bar{z}_i + \bar{B}_{yv.xz} \bar{v}_i + \bar{\mu}_i \quad (39)$$

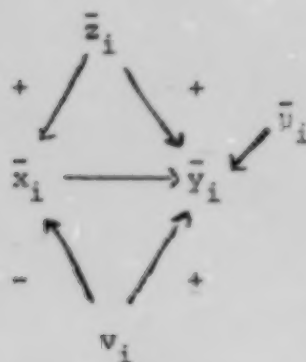
(See Figure 13.) The specification bias which arises from using (38) rather than (39) is:

$$E(\bar{b}_{yx}) - \bar{B}_{yx.zv} = \bar{B}_{yz.xv} \bar{b}_{zx} + \bar{B}_{yv.xz} \bar{b}_{vx} \quad (40)$$

The resulting aggregation bias is therefore:

$$\begin{aligned} E(\bar{b}_{yx}) - E(b_{yx}) &= B_{yx.zv} + B_{yz.xv} \bar{b}_{zx} + B_{yv.xz} \bar{b}_{vx} \\ &\quad - (B_{yx.zv} + B_{yz.xv} b_{zx} + B_{yv.xz} b_{vx}) \\ &= B_{yv.xz} (\bar{b}_{vx} - b_{vx}). \end{aligned} \quad (41)$$

Figure 13: Properly Specified Macromodel.



Since the specification bias in the micromodel is positive, aggregation gain requires that the aggregation bias be negative. In fact, this will always occur, since grouping values of x and y by v , an omitted variable, is equivalent to grouping by z . In other words, equation (41) is similar to equation (30); just as with grouping by z , grouping by v will increase the already negative relationship between x and y . Thus, $(\bar{b}_{yx} - b_{yx}) < 0$ while $B_{yv.zx} > 0$, so the whole term is negative. In sum, when data are grouped on a confounding variable omitted from the microrelation whose influence on that microrelation is opposite in sign and less than the influence of another omitted but confounding variable, aggregation gain is inevitable. If, however, the confounding influence of the excluded variables were in the same direction, or if data were aggregated on the confounding variable with the stronger influence, then grouping would generate aggregation loss.

Finally, there is one important instance in which there will always be an aggregation gain. Regardless of how the data are grouped, proper specification of the group level relation will always completely remove any effects of a poorly specified micromodel. Estimates of an individual level relation based on grouped data, when the group level relation is perfectly specified, will always be unbiased, even if the perfect microrelation is unknown or if the microdata are unavailable. Proof of this is shown in Table 1. Regardless of the grouping strategy, specification bias is zero for the aggregate level relation and measures the influence of the omitted confounding variable, z , at the individual level. Consequently, aggregation bias will always be the negative of specification bias in the microrelation, and data aggregation will cancel the effects of a poorly specified microrelation. The implications of this result are straightforward and important. To the extent that an aggregate relation is better specified than an individual relation, using aggregate data will result in estimates of individual level behavior which are less biased than estimates derived from individual level data. If aggregate data are of better quality and breadth than individual level data, so that more of the known variables can be explicitly and more carefully measured at the aggregate than at the individual level; or if a theory of group level behavior is better grounded than the corresponding

individual level theory, then using aggregate data will be superior to using individual level data. In short, individual level information and models of individual level behavior are no panacea. If proper specification at the group level is possible, reliance on individual level information is both unnecessary and costly.

IV.4 Summary

By considering together the issues of grouping and specification the distinct conditions under which aggregate level data will yield biased and unbiased estimates of individual level relations have been delineated. This approach to the problem of disaggregation also discloses the distinct situations in which the grouping of individuals produces aggregation gain and aggregation loss in parameter estimates.

Aggregation bias always results from the improper specification of group or individual level relations. When both these relations are correctly specified, the use of aggregate data will provide unbiased estimates of individual behavior regardless of whether the aggregate data are grouped by an independent variable, a dependent variable, or by both variables. The strategy of grouping by the dependent variable may, however, require reconstruction of the aggregate relation so that the confounding component of the microdisturbance created by these grouping procedures is included explicitly in the macrorelation. Respecification

Table 1: Specification and Aggregation Bias for
Perfectly Specified Macrorelations.

$$\text{Macro specification bias: } E(\bar{b}_{yx}) - \bar{B}_{yx.z} = 0$$

$$\text{Micro specification bias: } E(b_{yx}) - B_{yx.z} = B_{yz.x} b_{zx}$$

$$\begin{aligned} \text{Aggregation bias: } E(\bar{b}_{yx}) - E(b_{yx}) \\ &= B_{yx.z} - (B_{yx.z} + B_{yz.x} b_{zx}) \\ &= - B_{yz.x} b_{zx} \end{aligned}$$

of the macrorelation in this fashion ensures that it is specified properly and that no aggregation bias will be present. Alternatively, because of the practical difficulty of isolating the confounding component, it may be more useful to conceptualize the causal process as being reversed at the aggregate level when data are grouped by the dependent variable. Unbiased estimates of the effect of an individual level independent variable, x , on the individual level dependent variable, y , can then be computed from estimates of the effect of $\bar{y}$ on $\bar{x}$ at the aggregate level. In either case, proper specification ensures that aggregation bias vanishes.

In addition to the case in which both relations are specified properly, three other mutually exclusive and exhaustive cases were also considered: a properly specified micromodel and an improperly specified macromodel, an improperly specified micromodel and a properly specified macromodel, and an improperly specified micro and macromodel. When a microrelation is specified properly, but the macrorelation is not, aggregation bias, which is the difference between the expected values of the macro and micro regression coefficients, will be equal to the specification bias in the macrorelation. Thus aggregation loss will always occur. When the aggregate level relation is properly specified but the microrelation is not, aggregation bias will exactly cancel the specification bias in the micromodel. In this instance, the use of grouped data will always yield an aggregation gain, since

macrolevel parameters will be less biased than the corresponding individual level parameters.

Most complex is the situation in which both the micro and macrorelation are misspecified. In this instance, the use of aggregate data to estimate individual behavior may generate no aggregation bias or may generate aggregation bias that either amplifies or reduces specification bias in the micromodel. The actual result depends upon the grouping strategy, the true microrelation, and the misspecified micro and macrorelations. When the macrorelation is better specified than the microrelation there always will be an aggregation gain; aggregation bias will serve to diminish specification error in the microequation. Generally, the analysis does not assume that the macrorelation is better specified than the microrelation. Several distinct cases were separately considered: first, cases in which only one variable is excluded from the microrelation, and second, cases in which two or more variables are excluded from the microrelation. For the first set of cases, when data are grouped by an exogenous independent variable included in the misspecified microrelation no aggregation bias is created. When data are grouped by an endogenous independent variable included in the microrelation there will always be aggregation bias which tends to amplify the specification bias in the microrelation. Similarly, grouping by the variable which is excluded from the misspecified model will always amplify

specification error. However, grouping by a variable related to both an included and an excluded independent variable may either amplify or reduce specification bias.

Both aggregation gain and loss can also occur in the second set of cases in which two or more variables are excluded from the micromodel. If these two variables have opposite but unequal confounding effects, and if, in addition, the data are grouped by the excluded variable with the smaller confounding effect, aggregation bias will be in the opposite direction of specification bias in the micromodel and an aggregation gain will be achieved. If, however, data are grouped in this fashion and the confounding effects of the excluded variables are equal, or in the same direction, aggregation gain is impossible; only aggregation loss can occur.

IV.5 Implications

This study offers new approaches to the problem of inferring individual level behavior from data collected for aggregate units. It covers virtually every situation which the practicing policy analyst is likely to encounter, offering not a panacea for the disaggregation problem, but an analysis of the differences between expected values of statistical parameters at the individual and aggregate levels of measurement.

The methods employed are perfectly general, and can be extended to the analysis of any disaggregation problem.

The study identifies the situations in which it is theoretically possible to make valid cross-level inferences and sets forth the requirements that must be fulfilled in each instance. Indeed, only when certain types of contextual properties influence individual level relations is there an unbridgeable gap between the behavior of individuals and the study of aggregates. In all other cases, a properly constructed, macrolevel regression model can yield both unbiased and unique estimates of individual level relations.

The major innovations of this work are an integral treatment of the issues of specification and grouping, and a separate treatment of the problems of obtaining unbiased and unique estimates of individual level parameters from aggregate level data. The study demonstrates that aggregation bias in cross-level parameter estimates results from grouping effects that alter the proper specification of the aggregate level regression model. These effects make it necessary to include the grouping variable explicitly in the macrorelation; its inclusion may or may not be necessary in the microrelation. In every instance, aggregation bias is the difference between specification bias in the aggregate and individual level parameter. There will be no aggregation bias if regression equations at both levels of analysis are properly specified, or if they are both improperly specified with equal specification bias in the parameter being considered.

Since it is always theoretically possible (if not feasible) to specify properly both micro and macroequations, aggregation bias is not, in theory, an insurmountable problem. The study also sets forth the conditions under which aggregation bias will either increase or decrease the amount of specification bias in parameter estimates derived from an improperly specified individual level model. Moreover, the previous chapter demonstrates that aggregation bias is not the only difficulty created by the attempt to estimate individual level behavior from group level data. When the individual level model includes a contextual variable there is the additional problem of obtaining unique cross-level estimates of individual and contextual effects. The previous chapter distinguishes those contextual properties which cannot be separately estimated at the aggregate level from those which can be separately estimated through the use of multiplicative and nonlinear regression models.

These findings have important practical implications. They clearly demonstrate that for investigators concerned with obtaining unbiased estimates of individual level relations, data collected at the aggregate level is not necessarily less useful than data collected at the individual level. First, there can occur what might be termed a "pure" aggregation gain--a reduction in specification bias that results from the grouping process itself. Such aggregation gain can result from proximity grouping or from grouping strategies chosen by the investigator.

If the investigator seeks to minimize specification bias in parameter estimates and has the option to group a set of individual level data, the optimal strategy is to aggregate on a variable which is unrelated to the dependent variable and which tends to diminish the strength of the relationships between included and excluded independent variables. Second, and more importantly, policy analysts will often be able to construct better specified regression models for macro than for microlevel data; at the aggregate level they may be able to measure a wider range of relevant independent variables. Estimates taken from a well-specified aggregate level model may better reflect individual behavior than estimates taken from a poorly specified individual level model.

More broadly, the results suggest that investigators should focus primarily on the specification of micro and macro level regression models rather than on forms of grouping. The process of grouping individuals into units is important only as it influences the proper specification of the aggregate level model. To the extent that it is possible to eliminate specification bias in group level equations by explicitly including the grouping variables, it becomes less important to worry about the issue of grouping. When the aggregate level model is correctly specified, there will either be no aggregation bias in parameter estimates, or aggregation bias will serve to cancel specification bias in an improperly

specified individual level model. When the aggregate level model is not correctly specified, however, aggregation bias will equal the difference between specification bias in the micro and macromodels. Grouping on either the dependent variable or an excluded independent variable may contribute to differences in specification bias at the two levels of analysis. But the magnitude and direction of aggregation bias can be determined only from knowledge of the relations that characterize the properly specified individual level model.

It is also critical that policy analysts recognize the existence of contextual effects on individual behavior. In some cases it may be necessary to estimate such effects at the aggregate level through the use of nonlinear and multiplicative regression. In other cases, it will be impossible to distinguish between the operation of individual and contextual variables. The problem of contextual properties is particularly important for policy research. Contextual properties often represent factors which can be manipulated, whereas individual properties generally cannot. Unless contextual effects can be uniquely estimated it will be difficult to determine the implications of alternative policy decisions, as the next chapter clearly demonstrates.

V. SCHOOLS OR STUDENTS: A REANALYSIS
OF THE COLEMAN REPORT

After countless reanalyses and replications two fundamental conclusions of the Coleman Report (1966) remain widely accepted. Even staunch critics of the report concede both the crucial importance of family background for student achievement and the relative insignificance of such contextual properties of a student's environment as the characteristics of peers and the quality of teachers and school facilities. Since contextual properties can be manipulated, whereas the family background of individual students cannot, these conclusions have disturbed and frustrated those charged with shaping educational policy. If Coleman's findings are correct, such traditional educational reforms as improving school facilities, reassigning students, and upgrading the quality of teachers are likely to have little or no effect on the cognitive development of students.

Accurate empirical estimates of the impact of individual and contextual properties, however, depend critically on whether schools or students constitute the unit of analysis. Yet replications and reanalyses of the Coleman report disclose little agreement about the unit of analysis most appropriate for the study of student achievement. These inconsistencies in the choice of units can have important substantive implications. Previous chapters have firmly established that changes in the unit of analysis may generate substantial changes in the results of quantitative analysis. This investigation, a reanalysis of Coleman's data, demonstrates

significant disparities in identical models of student achievement estimated for students and schools. First, standardized measures like correlations and beta weights are particularly sensitive to changes in units of analysis. The correlation between family background and achievement increases considerably when schools rather than students constitute the unit on which data are measured. Family background is a good predictor of student achievement only when schools are the unit of analysis. Its poor performance when students are the unit of analysis suggests that very little is known about cognitive development (Hanushek and Kain, 1972: 130). Moreover, as measured by standardized regression coefficients, the impact of family background variables on achievement appears to be very large when schools are the unit of measurement. Yet much of this impact is an artifact of the grouping process and has no causal significance.

Second, when schools are the unit of analysis, quantitative techniques may be unable to separate the influence of contextual from individual variables. The ability to distinguish between the effects of individual and contextual properties depends on the contextual characteristics that influence a particular situation. Of the six characteristics identified in Chapter IV, three require the use of individual level data to isolate their unique impact on achievement. Of these, two appear to be effective determinants of student achievement: the family background and cognitive ability of a student's

peers. When investigators analyze schools rather than students, they are likely to attribute erroneously the contribution of these variables to the students themselves rather than to their environment. Manipulable properties of a student's environment, although important influences on student achievement, may be obscured by the analysis of aggregate data. Only by using students rather than schools as the unit of analysis can investigators avoid such faulty and misleading inferences.

Finally, the study shows that, under certain circumstances, using both individual and aggregate level data to estimate identical models also establishes a test for the adequate specification of the individual level model. As Chapter V demonstrated, adequate specification assures investigators that estimates of the impact of independent variables are unbiased. Thus, the use of dual levels of analysis is an important tool with which analysts can assess the internal validity of their research. For educational policy research, testing models of achievement using both students and schools can help assess the validity of estimates of the impact of non-manipulable student characteristics and manipulable contextual properties. Although this tool demands costly data on students, it significantly reduces the likelihood of misleading, and therefore even more costly, estimates.

V.1 The effects of grouping: standardized measures, contextual effects and the specification of models

Suppose that a student's achievement (Y) is affected by his race (X_1) and by his socioeconomic status (X_2).

Suppose also that the error term in the following regression equation is uncorrelated with X_1 and X_2 :

$$Y = b_1^* X_1 + b_2^* X_2 + e. \quad (1)$$

In this equation, b_1^* and b_2^* are beta weights. In lieu of data regarding the achievement, race and socioeconomic status of individual students, schoolwide means may be used. Instead of equation (1), the equation:

$$\bar{Y} = \bar{b}_1^* \bar{X}_1 + \bar{b}_2^* \bar{X}_2 + e \quad (2)$$

is used.

The magnitude of standardized measures associated with equation (2) relative to those associated with equation (1) depends on the grouping process. Several grouping processes could determine how students are allotted to schools. At one extreme, a student's race and socioeconomic status could dictate school attendance. Each school would then be more homogeneous with respect to race and status than the population of students in the district. Such an assignment procedure groups students according to their values on the two independent variables in equation (1). At the other extreme, a student's achievement score could determine which school he attends. In this case, high achievers attend schools with other high achievers; low achievers attend with other low achievers. Empirically, school assignment is more likely to approximate the first rather than the second extreme, especially for data collected in the mid-1960's. Thus, examination of the behavior of standardized measures of association when data have been grouped by values of X_1 and X_2 is instructive.

If students are grouped into schools according to their race and status, the magnitude of standardized measures of association increases. Presupposing that the model is properly specified, Chapter IV showed that grouping by independent variables makes $\bar{R}_{Y_1 X_1 X_2}^2$ exceed $R_{Y_1 X_1 X_2}^2$. Aggregate analysis may therefore suggest that more is known about the determinants of student achievement than individual analysis would warrant.

In addition, grouping by race and status increases the magnitude of beta weights, so that the absolute value of $\bar{b}_1^*$ and $\bar{b}_2^*$ will be larger than b_1^* and b_2^* respectively. Thus, as long as equation (1) is properly specified, and as long as schools are relatively homogeneous with respect to X_1 and X_2 , the standardized influence of race and status on achievement will appear larger when data are measured on schools instead of students. In contrast, under these same conditions, the unstandardized regression coefficients are stable from one unit of analysis to another. Analyses of student achievement that rely on aggregate data to compute standardized measures of influence are therefore likely to overestimate the effects of individual level variables like race and status.

Nevertheless, even if unstandardized rather than standardized coefficients are used, aggregate level data are likely to remain unsatisfactory for the estimation of contextual effects. In particular if either the socioeconomic status ($\bar{X}_2$) or ability ($\bar{Y}$) of peers directly affects individual achievement, obtaining estimates of their unique impact is impossible from aggregate data, as Chapter IV showed.

Even if contextual properties posed no aggregation problems for educational research, the use of aggregate data to study students could still be unwise. Chapter V showed that aggregate data produce unbiased estimates of individual behavior only when the aggregate level equations are properly specified. Some grouping procedures alter the specification of otherwise well-specified individual level models, while others do not. If attendance at a given school is largely determined by a student's race and socioeconomic status, then school level data approximate data grouped by the independent variable. Grouping by the independent variable does not alter the specification of the aggregate level equation. If the individual level model is properly specified, regression parameters estimated using data on students will be close to parameters estimated from data on schools.

Alternatively, students may be assigned to schools on the basis of their achievement scores. For example, communities with separate college preparatory and vocational high schools may approximate assignment by achievement score. Parents may also choose to live near schools whose achievement levels are similar to their children's achievement. In either case, school level data constitute data grouped by the dependent variable. Chapter V showed that grouping by the dependent variable alters the specification of aggregate relations by reversing the causal direction implicit in the individual level relation.

Finally, some combination of these extremes may characterize school level data. If so, then school level

data are grouped by both independent and dependent variables. Grouping by these variables also requires altering the aggregate level equation if it is to be properly specified. Statistical controls for whatever process causes students of similar race, status and achievement to attend the same school assure proper specification. Unfortunately, the underlying cause of this kind of grouping is rarely measureable or even identifiable in many practical applications.

Nevertheless, determining which grouping process characterizes school level data has important implications for educational research. If school level data approximate grouping by both independent and dependent variables, respecification is more a theoretical than practical possibility. It would probably be better to use student level data instead, despite additional costs. If school level data approximate dependent variable groupings, reversing the causal direction results in proper specification. Estimates of individual level parameters from aggregate data will then be unbiased, at least in the two and three-variable case, as long as estimates of the total or within group variance of each variable are available or can be approximated from other sources, as Chapter V showed. Finally, no alteration in the aggregate level model is necessary if school level data constitute grouping by the independent variable. Aggregate data would then yield unbiased estimates of individual and contextual effects.

It is possible to determine whether grouping by the independent variables best characterizes school level

data by simultaneously examining the behavior of regression estimates at the group and individual level. Only if the individual level model itself is well specified will the identical model estimated from data grouped by the independent variables also be correctly specified. If the individual level model is not properly specified, grouping by the independent variables results in an aggregate model that is also improperly specified. Therefore, only if an aggregate model is both well specified and uses data grouped by the independent variables will its estimated parameters equal estimates from the identical individual model. In other words, both conditions must be met when estimates from identical individual and aggregate models are stable.* If, on the

* There is one important exception to this rule. Stability can also be obtained if data are grouped by an independent variable, x , included in the individual relation, when two other criteria are met. First, the individual relation must be misspecified by the omission of a confounding variable, z , and second, x must be an exogenous variable. (For a demonstration, see Chapter IV.) In particular, the omitted variable z may represent the unknown mechanism that could make school level data approximate aggregation by the dependent as well as the independent variables. If, in addition, the independent variables in the individual relation are exogenous, stable estimates will be observed, but the estimates will be equally biased. However, the likelihood that the independent variables, race and socioeconomic status, are exogenous, seems small. Suppose, for example, that the grouping mechanism, z , is the income and racial makeup of the housing in the school's sending area. It is more plausible to suppose that the relation between x , the individual level properties, and z , the housing characteristics, is reciprocal than to suppose that x is exogenous. Only when x is exogenous will misspecified equations grouped by x yield estimates identical to estimates from ungrouped data.

other hand, one or both of these conditions are not met, the estimates will be unstable. However, it is impossible to tell which condition is violated.

The following results indicate that a model in which student achievement is determined by race and socioeconomic status and which incorporates peer status as a direct contextual effect, yields stable estimates. Predictably, stability is only observed when unstandardized rather than standardized measures are used. According to the foregoing logic, this model is well specified. School level data that approximate aggregation by race and socioeconomic status apparently produce unbiased estimates.

If unbiased estimation were the only desiderata for policy relevant educational research, use of school level data to study students would clearly be advisable. Valid estimation could be obtained at much less cost. However, policy research must be judged by its evaluative capacity. Unfortunately, the properly specified model also includes a direct contextual property that is defined in terms of an independent variable included in the individual level equation. Using aggregate data makes it impossible to obtain unique estimates of the impact of this type of contextual property. Yet analyzing the effects of changes in the socioeconomic status of those with whom a student attends school is of critical importance to educational policy. Therefore, even though aggregate data appear to yield unbiased estimates, their failure to produce unique estimates of contextual effects jeopardizes the utility of aggregate data for educational policy research.

V.2 Data source and measurement

The data for the analysis came from the 12th grade portion of Coleman's Equal Educational Opportunity Survey, completed in 1964. Of the 97,654 12th graders in Coleman's sample, a 10% random sample was selected for this analysis.* The variables measuring student characteristics, teacher quality, and school facilities came from the 12th grade student, teacher and principal questionnaires, respectively. Other measures of school characteristics were specially constructed for this analysis by aggregating the relevant properties of students attending the same school. The school level analysis has the same effective number of observations as the student level analysis. In other words, it weights each school by the number of students in the school.**

Since the purpose of this study is to investigate identical models at two different levels of analysis, each variable is comparably measured at each level. For students, verbal achievement is the difference between the number of correct and incorrect answers as a proportion of the total number of answers. In other words, it

* A sample of Coleman's original sample was used simply to reduce computer time.

** Coleman's survey oversampled schools with predominant minority proportions. However, it proved impossible to reconstruct these weights from publically available information. Armor (1972: 183) reports that his weighted and unweighted analyses produced very similar estimates of racial differences. Although means differed, differences between means did not. This analysis concerns only the latter; absolute mean values are not crucial to the study.

measures the relative predominance of right over wrong, or wrong over right, answers. Its value ranges from -1.00 to +1.00. For schools, verbal achievement is a function of the average number of correct, incorrect, and total answers given by 12th grade students attending the same high school. The aggregate measure of achievement also ranges from -1.00 to +1.00, and reflects the relative predominance of the average number of right to wrong answers within a school.

At the individual level, race is a dummy variable, where whites scored 1 and nonwhites, 0. At the school level, race is the proportion of students who are white.* The indicator of socioeconomic status is parental education. For students, the measure of parental education averages the mother's and father's education. At the school level, the measure weights the educational level of each parent, and then averages the score of each parent.** Thus, changes in the level of analysis do not change the scales of measurement for these variables.

* The regression coefficient associated with a dummy variable can be interpreted as the proportionate difference in achievement between whites and non-whites. Thus, even though the individual level indicator is interval, their regression coefficients have identical interpretations.

** The values of education ranged from 1 (some grade school) to 8 (attended graduate school). For the i -th student, education was $X_{ij} = (X_{mij} + X_{fij})/2$, where X_{mij} = educational level of mother of i -th student and X_{fij} = educational level of father of i -th student. For the j -th school, education was $\bar{X}_j = (\sum_i X_{mij} P_{ij} + \sum_i X_{fij} P_{ij})/2$.

Regardless of whether the variable is measured for students or schools, a unit change has the same meaning.

The measure of school facilities is more reflective of diversity and quantity than quality. The principal questionnaire included data on the number of books in the school library, the availability of science laboratories, the presence of guidance counselors, and the existence of a large variety of curricular offerings. Scores on these indicators were combined to produce a single index that reflects the variety of facilities potentially available to students in each school.

Like school facilities, teacher quality is a contextual variable defined without reference to variables included in the individual level model. In this study, teacher quality is measured by the average number of right over wrong answers as a proportion of the average number of all answers given by teachers on a verbal achievement test. Since it was impossible to match students to teachers, only schoolwide averages could be computed.*

where p_{ij} = proportion of students whose parents had educational level of X_{mij} or X_{fij} . Missing values were eliminated from all computations.

* Equations including multiple measures of teacher quality were originally estimated. Included as indicators were the percent of teachers who were white, the pupil-teacher ratio, the percent of teachers with a masters' degree or better, the percent of teachers whose parents had college degrees, full-time years of teaching, and the percent who were certified. Of these, only the teachers' verbal ability and the percent white had any significant

Each of these indicators is simple. Single indicators and additive indexes were chosen instead of the multiple indicators and complex indexes frequently used in other analyses of the Coleman data. For the particular purpose of this analysis, neither of the latter approaches is satisfactory. Using multiple indicators in a regression equation often results in very small partial coefficients, because indicators of the same concept tend to be highly correlated. Artificially small partials can be avoided using multiple partial correlation coefficients (Sullivan, 1971). However, this analysis depends on standardized (beta) and unstandardized (b) regression coefficients. Unfortunately, no multiple partial analogues of these measures exist. Finally, the use of complex indices raises another issue of aggregation that should be considered separately from those considered in this study. It is unclear whether each indicator should be combined into a single index before or after aggregation over students. In addition, the use of complex index construction techniques like factor analysis makes interpretation of unstandardized

impact on student achievement, regardless of the unit of analysis. The racial composition of teachers and their verbal ability are highly interrelated. Since both variables could not be included in a single equation, only verbal ability was retained due to its theoretical importance and policy significance.

Measuring verbal ability of teachers at the school rather than classroom level undoubtedly reduces its variance. The analysis to follow therefore probably underestimates the amount of variance shared between achievement and teacher quality.

regression coefficients unclear, particularly for policy analysis. It is meaningful to state that a percentage increase in an independent variable is associated with some percentage increase in the dependent variable. To claim that a unit increase in a factor score is associated with change in the dependent variable has dubious meaning in policy research.

V.3 Results

Table 1 shows the standardized (Beta) and unstandardized regression coefficients obtained from the regression of verbal achievement on race and parental education at the student and school level. At the individual level, family background factors explain only 26% of the variance in achievement. In contrast, at the aggregate level, these same factors explain 83% of the between-school variance in achievement. Analyses that rely on individual data thus suggest how little we know about the determinants of educational achievement. In contrast, analyses that rely on aggregate data attest to the importance of family background.

Such an inference from aggregate data is, however, an artifact. First, the variance of achievement is greater within than between schools, as others have noted (Hanushek and Kain, 1972: 129). Consequently, there is less variance to be explained at the aggregate than at the individual level. Second, evidence presented subsequently suggests that school level data approximate data grouped by both independent variables. Grouping by independent variables maximizes their variance. The high aggregate R^2 is attributable, then, to less variance in Y and greater variance in X_1 and X_2 than at the individual level. In short, although family background factors explain a high proportion of the variance in achievement between schools, the high aggregate R^2 is no evidence that an individual student's achievement depends on his family background. Instead, it suggests

Table 1

Regression of achievement on race and education:
students versus schools.

	Student		School	
	Race (X_1)	Education (X_2)	Race ($\bar{X}_1$)	Education ($\bar{X}_2$)
b	.004	.09	.004	.21
Std. error	.0002	.01	.0006	.003
Beta	.35	.31	.53	.56
R^2		.26		.83

that students who have similar racial and socioeconomic characteristics are likely to attend the same school.

Evidence in Table 1 is also consistent with the argument that school level data are grouped by family background characteristics. Because grouping by independent variables maximizes their variance, standardized measures of association like r and B_{ata} should increase with increasing levels of aggregation. We have already shown that the coefficient of multiple determination (R^2) increases. The simple correlations do also. At the individual level, the Pearson product-moment correlation between parental education and achievement is .37; at the aggregate level, it is .77. The individual level correlation between race and achievement is .40; the aggregate correlation is .75. The correlation between the two independent variables also increases, but not dramatically enough to make collinearity a threat to the aggregate level multiple regression equation. The correlation between students' race and parental education is .16; for schools, the correlation is .38.

Table 1 also shows that, as measured by standardized partial regression coefficients (B_{etas}), the influence of race and parental education on achievement appears about 60% greater at the aggregate than at the individual level. This effect is consistent with the hypothesis that data on schools constitute data grouped by family background.

Moreover, if the student level equations are properly specified and if data are grouped by independent variables, unstandardized regression coefficients should remain unchanged. Table 1 shows that the coefficient for race is stable; the education coefficient, however, is larger at the aggregate than the disaggregate level.

Rather than reject the hypothesis that school level data are grouped by family background, respecification of the student level equation may yield better results. Inclusion of the parental education of students' peers, represented by equation (18) in Chapter III, does indeed yield stable regression weights. (See Table 2.) In the absence of individual data, the aggregate version of such a model, represented by equation (22) in Chapter III, is indistinguishable from equation (24) in Chapter III, the aggregate version of a model that omits the family background of a student's peers. However, equation (22) in Chapter III, shows that the individual level coefficient associated with a student's own parents education plus the coefficient associated with his peer's parental education should equal the single aggregate level coefficient associated with parental education. Table 2 shows that the empirical results very nearly approximate theoretical expectations.

The sum of the two parental education coefficients estimated from individual level data ($.08 + .09 = .17$) is very close to the single education coefficient estimated from aggregate level data (.21).

In sum, the stability of the regression coefficients in Table 2 indicates two important results. First, the individual level model that generated those estimates, represented by equation (2), is probably reasonably well-specified. Second, the school level data approximate data grouped by family background and not by achievement. Both of these conditions must be met when individual and aggregate level regression coefficients are stable.

These findings have important implications. First, under ordinary conditions, when a model is well-specified and data are grouped by the independent variables, no aggregation bias results. In other words, unstandardized regression estimates are unbiased, regardless of whether costly individual level or less expensive aggregate level data are used. However, the apparent cost-effectiveness of aggregate data to study student achievement is specious in this case. The best specified model includes a contextual variable that has direct effects on achievement and is a linear function of an individual level independent variable. Under these circumstances, use of aggregate data cannot reveal the separate impact of contextual and individual factors.

Although the addition of peer parental education to the original model does not increase explained variance, it does result in a better specified individual level model. Moreover, only individual level data disclose that the effects of peer and individual parental education are about equal, as measured by unstandardized

Table 2

Regression of achievement on race, parental education,
and peer parental education: students versus schools.

	Student			School	
	Race (X_1)	Education (X_2)	Peer educ. ($\bar{X}_2$)	Race ($\bar{X}_1$)	Education ($\bar{X}_2$)
b	.003	.08	.09	.004	.21
Std. error	.0002	.005	.01	.0006	.003
Beta	.33	.29	.13	.53	.56
R^2		.27			.83

partial regression coefficients. Increasing the average educational level of the parents of those with whom the average student attends school will increase his verbal achievement by 9%, according to Table 2. Reliance on aggregate level data makes it impossible to tell what portion of the single aggregate level coefficient is attributable to contextual factors. In addition, reliance on aggregate data makes it impossible even to distinguish between a model which includes contextual variables of this sort and one which does not.

A better specified model, of course, might show interactive rather than direct effects of peer socio-economic status. (See equations (20) and (21) in Chapter III.) If so, then using aggregate data would be a cost-effective strategy for studying achievement. However, the instability of the regression coefficients estimated from student and school level data suggests that the direct effects model is superior. (See Tables 3 and 4.) Only the coefficients for race are stable from one unit of analysis to another. Since the results from Table 2 suggested that the school level data are grouped by race and parental education, instability is accounted for by poor specification.

Although the parental education of peers seems to have direct contextual effects on achievement, school facilities appear to have no effects. In fact, Tables 5, 6 and 7 are consistent with previous statements about the seeming irrelevance of gross measures of the quality of school facilities for student achievement. No matter

Table 3

Regression of achievement on race, parental education and parental education times race: students versus schools.

	Students			Schools		
	Race (X_1)	Educ. (X_2)	Educ. x Race ($\bar{X}_2\bar{X}_1$)	Race ($\bar{X}_1$)	Educ. ($\bar{X}_2$)	Educ. x Race ($\bar{X}_1\bar{X}_2$)
b	.006	.09	.0006	.003	.19	.0001
Std. error	.0008	.005	.0003	.0003	.008	.0001
Beta	.54	.31	.19	.49	.55	.05
R^2			.26		.83	

Table 4

Regression of achievement on race, parental education and parental education squared: students versus schools.

	Student			Schools		
	Race (X_1)	Educ. (X_2)	Educ. x Race ($\bar{X}_2\bar{X}_2$)	Race ($\bar{X}_1$)	Educ. ($\bar{X}_2$)	Fac. x Race ($\bar{X}_2\bar{X}_2$)
b	.003	.01	.02	.004	-.20	.06
Std. error	.0002	.01	.003	.00006	.02	.003
Beta	.33	.04	.30	.55	-.59	1.15
R^2		.27			.85	

Table 5

Regression of achievement on race, parental education, peer parental education, and facilities: students versus schools

	Students				School		
	Race ($\bar{X}_1$)	Educ. ($\bar{X}_2$)	Peer Educ. ($\bar{X}_2$)	Facil- ities ($\bar{Z}$)	Race ($\bar{X}_1$)	Educ. ($\bar{X}_2$)	Facil- ities ($\bar{Z}$)
b	.003	.08	.09	.001	.003	.20	.003
Std. error	.0002	.005	.01	.002	.00006	.003	.0004
Beta	.32	.29	.13	.01	.52	.56	.06
R^2		.27				.84	

Table 6

Regression of achievement on race, parental education, peer parental education, and facilities times race: students versus schools.

	Students				Schools		
	Race ($\bar{X}_1$)	Educ. ($\bar{X}_2$)	Peer Educ. ($\bar{X}_2$)	Fac. x Race ($\bar{X}_1 \bar{Z}$)	Race ($\bar{X}_1$)	Educ. ($\bar{X}_2$)	Fac. x Race ($\bar{X}_1 \bar{Z}$)
b	.004	.08	.09	.001	.003	.20	.00003
Std. error	.0009	.005	.01	.003	.0002	.003	.00001
Beta	.34	.29	.13	.02	.40	.57	.14
R^2		.27				.84	

Table 7

Regression of achievement on race, parental education, peer
parental education, and facilities times parental education:
students versus schools

	Race (X_1)	Educ. (X_2)	Peer Educ. (X_2)	Fac. x Educ. ($\bar{Z}X_2$)	Race ($\bar{X}_1$)	Educ. ($\bar{X}_2$)	Fac. x Educ. ($\bar{Z}X_2$)
b	.003	.08	.09	.0002	.003	.18	.0006
Std. error	.0002	.01	.01	.0004	.00006	.005	.0001
Beta	.33	.27	.13	.02	.53	.51	.07
R^2		.26				.84	

whether school facilities are hypothesized to have direct or interactive effects, adding the indicator of facilities to the basic model reveals the inefficacy of hardware. The substantive or statistical insignificance of the unstandardized regression coefficients and the small Betas associated with the measure of facilities support such a conclusion. Table 5 shows that adding facilities to a model that includes peer parental education does not alter its specification. The stable regression coefficients indicate that a model which includes both peer status and facilities as direct contextual effects is reasonably well specified. Consequently, in the case of contextual variables like facilities which are defined independently of other variables, conclusions reached from aggregate and disaggregate data will be similar. Regardless of the unit of analysis, school facilities appear to have no impact on achievement.

Investigators have also hypothesized that the teacher's verbal ability has an important impact on the ability of the students he teaches. If the teacher's verbal ability has a direct effect, adding that variable to the basic regression which includes race, parental education and peer parental education does not significantly increase the explained variance, as Table 8 reveals. However, as measured by the unstandardized partial regression coefficient, it does have some statistical and substantive significance. A percentage increase in the teacher's score is associated with a 14% increase in the average students' verbal score. At the

Table 8

Regression of achievement on race, parental education, peer
parental education, and teachers' score: students versus
schools.

	Race (X_1)	Educ. (X_2)	Peer Educ. ($\bar{X}_2$)	Teacher Score ($\bar{Z}$)	Race ($\bar{X}_1$)	Educ. ($\bar{X}_2$)	Teacher Score ($\bar{Z}$)
b	.003	.08	.08	.14	.003	.17	.35
Std. error	.0002	.005	.01	.07	.00007	.003	.02
Beta	.32	.28	.11	.05	.41	.50	.23
R^2		.27				.86	

aggregate level, the corresponding coefficient is larger. While this might indicate either that the student level equation is improperly specified or that it is incorrect to assume that the student level data are grouped by race and parental education, no final conclusion seems warranted in this case. Failure to measure the verbal ability of the teacher in each of the student's classes could introduce significant error into the analysis. Moreover, at the aggregate level, the correlation between between race and the teacher's score is .64; instable estimates could, in this instance, result from collinearity.

If the direct effects model of teacher ability seems inconclusive at best, an indirect effects model appears unwarranted. The unstable coefficients in Tables 9 and 10 suggest that including the multiplicative terms unfavorably alters the specification of the rest of the model. The coefficients associated with race are unstable in Table 9. In Table 10, the sum of the two education coefficients for the student model do not sum to the coefficient for education in the aggregate model. Yet previously these coefficients were relatively constant from one unit of analysis to another.

Overall, inclusion of contextual variables that are defined independently of variables in the individual level relation poses no problem for the use of aggregate data to study individuals. Of these kinds of contextual variables, only the teacher's verbal ability appears to have any impact on student achievement. In contrast, it

Table 9

Regression of achievement on race, parental education, peer
parental education, and teacher score times race: students
versus schools.

	Students				Schools		
	Race (X_1)	Educ. (X_2)	Peer Educ. ($\bar{X}_2$)	Teacher Score $\bar{x}$ race (ZX_1)	Race ($\bar{X}_1$)	Educ. ($\bar{X}_2$)	Teacher Score $\bar{x}$ race (ZX_1)
b	.005	.08	.08	.23	.001	.18	.004
Std. error	.0006	.005	.01	.08	.0002	.003	.0003
Beta	.46	.28	.12	.14	.11	.52	.46
R^2		.28				.85	

Table 10

Regression of achievement on race, parental education, peer
parental education, and teacher score times education;
students versus schools

	Students				Schools		
	Race (X_1)	Educ. (X_2)	Peer Educ. ($\bar{X}_2$)	Teacher Score $\times$ Educ. ($\bar{Z}X_2$)	Race ($\bar{X}_1$)	Educ. ($\bar{X}_2$)	Teacher Score $\times$ Educ. ($\bar{Z}X_2$)
b	.003	.06	.08	.04	.003	.08	.13
Std. error	.002	.01	.01	.02	.0001	.005	.005
Beta	.32	.19	.11	.11	.40	.22	.45
R^2			.28			.87	

is clear that peer parental education, a contextual property derived directly from an independent variable included in the individual level model, has a direct effect on achievement. Unique estimation of this effect is impossible without data on individuals.

Finally, the ability of peers is a contextual property derived directly from the dependent variable in the individual level model. Any impact that it might have on achievement could only be detected with individual level data. Unfortunately, problems of collinearity as well as problems of aggregation make it impossible to estimate its impact with clarity. If peer ability has direct effects on individual ability, Table 11 shows that its impact is both statistically and substantively significant. A 1% increase in the achievement scores of peers increases the score of the average student by 36%, if the evidence is plausible. Oddly, however, the effect of peers' parental education is sizeably reduced from its estimated coefficient of .09 in Table 2. In Table 11, it is not even statistically significant. This result, however, simply reflects the very high correlation of .77 between peer ability and peer parental education. The presence of such high collinearity makes reliable estimation impossible.

Aggregate level data in this case are no help. The school level estimates in Table 11 are identical to those in Tables 1 and 2. In other words, a model based on direct effects of peer ability is indistinguishable from one in which peer parental education has direct effects as well as from one in which there are no contextual

Table 11

Regression of achievement on race, parental education,
peer parental education, and peer ability: students
versus schools

	Students				Schools	
	Race (X_1)	Educ. (X_2)	Peer Educ. ($\bar{X}_2$)	Peer Ability ($\bar{Y}$)	Race ($\bar{X}_1$)	Educ. ($\bar{X}_2$)
b	.003	.08	.002	.36	.004	.20
Std. error	.0002	.005	.02	.06	.00006	.003
Beta	.28	.28	.002	.18	.53	.53
R^2		.28				.83

effects at all. The partial regression coefficient of .20 associated with education in the aggregate model cannot be decomposed. It is impossible to tell whether its value reflects the combined effects of contextual and individual parental education, the impact of individual background weighted by the contextual effects of peer ability, some combination of these, or perhaps none of these at all.

It is difficult to evaluate a model incorporating indirect effects of peer ability. Assuming that data grouped by school approximate data grouped by family background, then stable individual and aggregate coefficients are a necessary indicator of adequate specification. But no aggregate level model can be estimated at all if peer effects are indirect, as equation (35) in Ch. III demonstrates. The coefficients associated with the individual level multiplicative term in Tables 12 and 13 are statistically significant. Moreover, the R^2 's are similar to those computed for a direct effects model.

In Table 13, the regression coefficient associated with peer parental education is not significant, in contrast to Table 12. However, this is an artifact of multicollinearity, since the correlation between peer parental education and the multiplicative term is .72. In short, both models of the indirect effects of peer ability are difficult to evaluate without the additional information gained by a comparison of individual and aggregate models.

Table 12

Regression of achievement on race, parental education,
peer parental education, and race times peer ability:
students versus schools

	Students				Schools
	Race (X_1)	Educ. (X_2)	Peer Educ. ($\bar{X}_2$)	Ability x_{race} (YX_1)	
b	.003	.08	.08	.17	
Std. error	.0002	.005	.01	.08	No estimates possible
Beta	.32	.29	.11	.04	
R^2			.27		

Table 13

Regression of achievement on race, parental education,
peer parental education, and education times peer
ability: students versus schools

	Students				Schools
	Race (X_1)	Educ. (X_2)	Peer Educ. ($\bar{X}_2$)	Ability $x_{\text{educ.}}$ (YX_2)	
b	.003	.08	.02	.07	
Std. error	.0002	.006	.02	.01	No estimates possible
Beta	.30	.25	.03	.15	
R^2			.28		

V.4 Summary and Implications

The study of student achievement is critical to the development of educational policy. Determining whether improvements in school facilities, upgrading teacher quality, or reassigning students has a direct impact on the performance of students is necessary if funds for education are to be wisely allocated. Such a determination requires the impact of these policy alternatives to be evaluated apart from the effects of characteristics of students that policy cannot affect. Analysts concerned with empirically estimating the independent impact of manipulable variables must choose between using students or schools as the unit of analysis. Gathering data on students, however, is very costly. Moreover, both parents and school administrators are increasingly unwilling to collect information from students. If they are willing to collect information, they are increasingly unwilling to release it. In contrast, gathering data on schools is cheaper and easier to obtain. It is therefore important to consider the implications of using data on schools to evaluate how potential policy changes affect students.

In considering these implications, this study compared several identical models of student achievement at two levels of analysis. Each comparison disclosed how exclusive reliance on aggregate data can often produce misleading results. Aggregate data can safely be used when data are grouped by independent variables, aggregate equations are properly specified, unstandardized estimates are used, and when the operative contextual

variables permit unique estimation. However, the absence of individual level data makes it impossible to determine whether these conditions are in fact met. In addition, simultaneous analysis of data from both levels can be used to test whether aggregate models meet the first two of these conditions.

The findings disclose that reliance on standardized measures estimated from aggregate data can be hazardous. More specifically, aggregate correlations and beta weights are inflated. Beta weights obtained from aggregate data tend to exceed weights obtained from individual level data. Moreover, the inflation is greatest for coefficients associated with variables on which the individual level data have been grouped. Students appear to be grouped into schools on the basis of their race and socioeconomic status. Grouping, in effect, maximizes the variance of race and status. Standardized measures of association and influence are sensitive to changes in relative variance produced by grouping. Consequently, school level estimates of the standardized influence of race and status on achievement are much larger than similar estimates based on student level data. In short, the importance attributed to family background may be exaggerated when standardized estimates from aggregate data are the source of information.

Unfortunately, substituting unstandardized for standardized aggregate estimates is often an unsatisfactory solution. Under certain circumstances, it is impossible to obtain unique estimates of the effects of contextual variables from aggregate data. Of six types

of contextual properties identified in the analysis, three cannot be assessed from aggregate data. This dilemma would be academic if individual level analysis revealed that these properties were unimportant. However, individual level analysis discloses that the socioeconomic status and ability of peers have direct effects on the achievement of students. The former is a contextual property derived from an independent variable included in the individual level relation. The latter is derived from the dependent variable in the individual level relation. Unfortunately, using aggregate data makes it impossible to estimate uniquely the effects of these types of contextual properties. Unique estimates must be obtained from data on students. Even though data on students are costly, unique estimates are essential for educational research to have policy relevance.

In addition to uniqueness, estimates of the impact of contextual variables should also be unbiased. Simultaneous analysis of individual and grouped data can help to assess the bias in individual level estimates. If aggregate data are grouped by the independent variables in an individual level model, and if both individual and aggregate models are properly specified, unstandardized regression estimates from grouped and individual data will be equal within sampling error. On the other hand, if estimates are not equal, one or both of the conditions does not apply; it is impossible to determine which.

Analysis of student achievement using data on both schools and students revealed one model that exhibited

stable coefficients. That model included race, parental education, and peer parental education as independent variables. The stable regression coefficients disclosed not only that this model is properly specified, but also that students are grouped into schools on the basis of their race and their socioeconomic status. Thus, while multilevel analysis cannot rule out the possibility that other models are also properly specified, it increases confidence in a model that includes the socioeconomic status of peers as a contextual variable.

In sum, despite additional costs, exclusive reliance on aggregate data diminishes an analysts' ability to evaluate empirically alternative models of student achievement. Distinguishing among these models is critically important if research on student achievement is to be relevant to educational policy. Proponents of bussing allege, for example, that the family background of a student's peers has salutary effects on achievement, while protagonists protest its injurious influence. Advocates of tracking assume that peer ability abets individual improvement. And people of various political persuasions aver that achievement is largely determined by the family; public policy, they conclude, can do little. This analysis reveals that when schools rather than students are the unit of analysis, each of these positions is reinforced by statistical evidence. Yet it is impossible for all of these positions to be correct simultaneously. Indeed, statistical evidence will only confuse rather than enlighten and inform as long as schools are studied instead of students.

VI. THE PROBLEM OF CROSS-LEVEL INFERENCE:
IMPLICATIONS FOR EDUCATIONAL RESEARCH

Educational research confronts investigators with problems of both aggregation and disaggregation. Data aggregation is necessitated by two circumstances. First, when individual level data are available but too detailed for the rapid perusal frequently demanded by decision-makers, the information must be summarized or aggregated. Second, when individual level data are desirable but too costly, data must be collected at the group or aggregate level. In both cases, the user wants to lose as little information as possible to maximize the utility of aggregate data for making inferences about individual level behavior. This study has considered alternative ways of both summarizing and grouping individual level data. Thus, the first purpose of this summary chapter is to suggest which, if any, of these alternatives are likely to be most useful in two areas of educational research: studies of student achievement, and studies of the fiscal impact of educational aid.

The disaggregation of educational data is similarly required when users of educational data must rely on summary or aggregate information, yet intend to make inferences about individual level behavior. Detailed information about individuals is frequently too costly or

time consuming to obtain. Under these circumstances, the procedures used to analyze aggregate data should produce valid inferences about individual behavior. This study has explored several such procedures in some detail. Thus the second purpose of this chapter is to suggest which, if any, of these are most useful in the areas of educational research mentioned above.

There are two strategies of data aggregation. Data can be aggregated into units, and then such parameters as group means, medians, can be analyzed. This strategy may be termed aggregation before estimation. Alternatively, the individual level information itself can be analyzed and the results of the analysis can then be aggregated. In other words, estimation may also precede aggregation. Those strategies based on estimation before aggregation are likely to be of limited utility in educational research, because they are restricted to time series data. While time series are available in educational finance data, the questionable comparability of observations from one time to another and the large number of variables for which information is unavailable mean that time series applications will be infrequent. Similarly, data on student achievement and related variables measured over time are, at present, usually unavailable.

Strategies based on estimation before aggregation may potentially be applied to cross-sectional data when individual units systematically affect one another's behavior. However, even this modification is likely to be

of limited applicability until theories predicting specific patterns of spatial influence are developed. Finally, estimating relations before aggregating them requires a difficult transformation when aggregate dependent variables have a direct influence on any individual level variable included in the analysis. Studies of the fiscal impact of intergovernmental transfers of education monies are unlikely to require the transformation. Studies of student achievement, however, necessitate the modification because the overall level of student achievement affects individual achievement, attitudes, and aspirations.

The complexity of procedures for aggregating estimates of individual level relations is also their virtue. The complexity, in effect, preserves information. In particular, contextual as well as individual level effects can be uncovered, as illustrated by the empirical investigation of the expenditure effects of federal aid. Without directly examining detailed information about each of 48 states, that study showed how states respond differentially to changes in the total amount of federal aid to education. More simple summarizing strategies would fail to produce this additional information.

Aggregation before estimation entails the choice of a way to group individual level data such that bias and unreliability in estimates of individual behavior are minimized. Grouping by the independent variable is

ordinarily considered to be the "optimal" grouping technique, according to these criteria. Several other criteria, however, should enter into the choice of a grouping procedure. First, grouping by the independent variable will yield unbiased estimates of individual level behavior only if the individual level model is correctly specified. To the extent that the analyst is uncertain about the determinants of individual behavior, grouping by the independent variable may only increase the bias in estimates of individual level relations. Second, even if the individual level relation is properly specified, the aggregate data likely to be of greatest utility to decision-makers are data grouped by administrative unit. Grouping by geographic proximity will, in general, result in biased estimates of individual behavior. Yet to the extent that proximity grouping approaches grouping by an independent variable, no greater bias or instability in estimates will arise than if data were grouped by the independent variable itself. Apparently, studies of the fiscal impact of state aid may approximate this situation. Similar studies can therefore rely on state level data to study school district fiscal behavior. Of course, even when proximity grouping does not approximate grouping by an independent variable, it can still yield unbiased estimates of individual level behavior. If the processes which make people gather in the same geographic areas are known, measureable and included in an otherwise properly specified aggregate level model, proximity grouping will always

result in unbiased estimation. Studies of the role of income, public service preferences, racial tensions, and place of employment in influencing residential choice may be of some use in this regard. An understanding of why people choose similar residential locations could permit the use of proximity grouping in many areas of educational research to reduce the costs of data with no reduction in validity or utility.

The choice of an optimal grouping strategy should also consider problems posed by contextual effects and standardized measures of association. While contextual effects do not appear in studies of the fiscal impact of state aid, they do appear in studies of student achievement. Moreover, the contextual effects that do emerge are those which pose the most intractable problems for the use of any kind of group level data. Specifically, the empirical analysis of student achievement makes it clear that both the achievement and status of peers influence student achievement. Regardless of how data are grouped, when these properties are present it is impossible to derive separate estimates of individual and contextual effects. While public policy cannot change the characteristics of individual students, it can change the context in which students learn. Assessing the effectiveness of any such change requires separate estimates of contextual effects. Yet, in studies of student achievement,

such estimates cannot be derived from aggregate level data. In these studies the dilemma posed by contextual effects may be so severe that any form of grouping is unreasonable.

Finally, all aggregate data, regardless of how they are grouped, will yield biased estimates of individual level behavior when unstandardized measures like correlations and Beta weights are used. To guide the formulation of educational policy, research must compare either the influence of alternative policies or the influence of manipulable relative to unmanipulable variables on some behavior of interest. Unless variables are measured in common units, standardized measures must be used to make these comparisons. Relevant variables are frequently measured in common dollar units in studies of educational finance, so that unstandardized measures of relative influence will provide the necessary information. In contrast, studies of student achievement rarely include dollar-based variables. The need to rely on unstandardized measures of relative influence in macrolevel studies of student achievement, in addition to the problem of contextual effects, suggests another reason for abjuring the use of grouped data.

Ecological regression is clearly the most fruitful technique for making inferences about individual level behavior from aggregate data. Ecological regression produces unbiased estimates of individual level behavior

when three conditions are met simultaneously. First, the ecological regression must be perfectly specified. This means that, under most conditions, the grouping variables must be included in the ecological regression. If the ecological groupings are equivalent to the groups that would be obtained had data been grouped on the dependent variable, the regression of the independent on the dependent variable must be found. These techniques help to ensure that the ecological regression is perfectly specified. Second, the analyst must rely on unstandardized measures of influence. To compare the relative influence of independent variables, he must use variables measured in common units. Third, ecological regression permits unique estimation of individual and contextual effects only when indirect behavioral effects or "pure" contextual effects are at issue.

The implications of these conditions for studies of educational finance favor the use of ecological regression in preference to analysis of individual level data. First, the empirical example suggested that, for studies of the fiscal impact of state aid, school district data grouped by state approximate the groups that would be produced had the districts been arrayed by state aid per pupil. Contextual effects pose no problems in this area of educational research. Moreover, variables are frequently measured in

common dollar terms, so that statements about the relative influence of independent variables can be based on unstandardized coefficients.

Most problematic for the use of ecological regression in fiscal studies is the necessity of proper specification. Models could be better specified if better indicators of the supply of educational resources and the demand for educational services were available. On the supply side, better indicators of aid reaching school children which is not funded through the U.S. Office of Education would be useful. For example, states and communities that provide generous matching funds in federally aided health, food, housing and public assistance programs may be using resources that might otherwise be available to schools. On the demand side, more detailed descriptions of pupil populations would be desirable in addition to specific data on enrollment in non-public schools. Equity considerations could be more directly addressed if data on tax effort were comparable from one unit of analysis to another and if information concerning the allocation of resources to specific educational programs were available. This information would go far towards assuring the proper specification of ecological models of fiscal impact. Such information, gathered at the state level, would be cheaper to obtain than if it were also gathered for all school districts. Moreover, this reduction in cost would

entail little or no sacrifice in the validity of analytical results.

Thus, ecological regressions are probably a less costly but equally robust means of studying the fiscal behavior of school districts than individual level regressions. It should be apparent, however, that individual district data may still be necessary for other purposes. In fact, the President's Commission on School Finance recommended that data be available not just at the district but also at the school level so that evaluators at all levels of educational administration can assess resource inequities in their own jurisdiction. Thus, the argument that state level data suffice for the study of fiscal impact is consistent with arguments that data measured at lower levels of aggregation are still necessary for other equally important purposes.

Ecological regression is probably an unsatisfactory technique for the study of student achievement, largely because contextual effects so importantly influence the performance of students. Even when interest centers on "pure" contextual effects (for example, the effects of school facilities, teacher training, and curriculum variations), these effects may be mediated by other contextual properties which must be measured. If unique estimates of these latter contextual effects are not of interest, then aggregate data will suffice. In general,

it is essential to isolate contextual from individual effects if educational research is to be relevant for policy planning. As a result, ecological data should not be used for the study of student achievement.

Two issues need further study if aggregate data are to be an acceptable substitute for individual level information. First, proper specification of macrolevel regression models frequently requires inclusion of the grouping variable. Given the utility of proximity grouping to policy research, it is necessary to specify what variables cause similar people to choose similar residential locations. Understanding the processes by which spatial groups are formed makes proper specification of ecologically based macromodels feasible.

Second, the choice between using aggregate and individual level data should also consider the level of analysis at which proper specification is most feasible. Sometimes it is more difficult to specify fully a model

at the individual than at the aggregate level; theories of aggregate behavior may be better elaborated and relevant individual level data may be unavailable. In these cases, use of aggregate data may yield more valid inferences about individual level behavior than would use of individual level data. Moreover, under two circumstances, the grouping process itself can compensate for a poorly specified micromodel, even if the macromodel itself is also imperfectly specified. When data are grouped so as to decrease the correlation between omitted and included independent variables, aggregation bias will compensate for specification bias. Aggregation gain can also result when data are grouped on an omitted confounding variable which has smaller absolute influence in an opposite direction than another omitted confounding variable. Ordinarily, of course, if we knew what these confounding variables were, they would not have been omitted. However, sometimes data are simply unobtainable. Practical applications of these grouping strategies do not appear imminent. Yet they are worth additional consideration because of the frequency with which educational researchers must rely on inadequately specified models.

These conclusions are based on the study of only one problem in educational research--that of aggregation and disaggregation. Other substantive and methodological concerns may dictate different conclusions. For example,

many issues involving the equity with which educational resources, services and outcomes are distributed necessitate the use of disaggregate observations measured at the district, school, classroom or student level. Similarly, reliance on student-level data in studies of achievement raises additional questions about confidentiality of information and non-independence of observations.

Answers to these questions may suggest the advisability of observations measured at more aggregated levels. Thus, the conclusions from this study are no panacea for all problems in educational research. Nevertheless, the study has clarified the utility, importance and methodology of properly specifying ecological models. Regardless of the reason for relying on aggregate data to study individual behavior, these procedures will increase the likelihood of valid ecological inference.

BIBLIOGRAPHY

1. Allen, R. D. G. Mathematical Economics. London: Macmillan and Company, Ltd., 1956: Ch. 20.
2. Armor, David J. "Equality of Educational Opportunity: The Basic Findings Reconsidered." Ch. 5 in Frederick Mosteller and Daniel P. Moynihan, eds., On Equality of Educational Opportunity. New York: Random House, 1972.
3. Bahl, R., and Saunders, R. "Determinants of Change in State and Local Government Expenditures." National Tax Journal 18 (1965): 50-57.
4. Barton, A. H. "Bringing Society Back In: Survey Research and Macro-Methodology." American Behavioral Scientist 12 (Nov.-Dec. 1968): 1-9.
5. Berke, Joel S.; Bailey, S. K.; Campbell, A. K.; and Sacks, S. Federal Aid: Who Benefits? New York: The Ford Foundation, 1973.
6. Blalock, Jr., Hubert M. Causal Inferences in Non-Experimental Research. Chapel Hill: University of North Carolina Press, 1964.
7. Blau, Peter M. "Structural Effects." American Sociological Review 25 (1960): 178-193.
8. Boot, J. C. G., and deWit, G. M. "Investment Demand: An Empirical Contribution of the Aggregation Problem." International Economic Review 1 (Jan. 1960): 3-30.
9. Break, George F. Intergovernmental Fiscal Relations in the United States. Washington, D.C.: The Brookings Institution, 1967.
10. Burkhead, Jesse; Fox, T.; and Holland, J. W. Input and Output in Large City High Schools. Syracuse: Syracuse University Press, 1967.
11. Burnham, Walter Dean. "Theory and Voting Research: Some Reflections on Converse's Change in the American Electorate." American Political Science Review 68 (Sept. 1974): 1002-10023.

12. Burstein, Leigh. "Issues Concerning Inferences from Grouped Observations." Paper presented at the Annual Meeting of the American Educational Research Assn., April 1974, Chicago, Ill.
13. Campbell, Alan K., and Sacks, Seymour. Metropolitan America: Fiscal Patterns and Governmental Systems. New York: The Free Press, 1967.
14. Campbell, E. Q., and Alexander, C. N. "Structural Effects and Interpersonal Relationships." American Journal of Sociology 71 (Nov 1965): 284-289.
15. Christ, Carl F. Econometric Models and Methods. New York: John Wiley, 1966.
16. Coleman, James S., et al. Equality of Educational Opportunity. Washinton, D.C.: U.S. Government Printing Office, 1966.
17. Crain, Robert, The Politics of School Desegregation. Garden City: Doubleday and Co., Inc., 1969.
18. Cramer, J. S. "Efficient Grouping: Regression and Correlation in Engel Curvel Analysis." Journal of The American Statistical Association 59 (March 1964): 233-250.
19. Crecine, John P. "A Simulation of Municipal Budgeting: The Impact of Problem Environment." In Ira Sharkansky, ed. Policy Analysis in Political Science, Chicago: Markham Publishing Co., 1970.
20. Davis, James A.; Spaeth, Joe L.; and Huson, Carolyn. "A Technique for Analyzing the Effects of Group Composition." American Sociological Review 26 (1961): 215-226.
21. Draper, N. R., and Smith, H. Applied Regression Analysis. New York: John Wiley and Sons, 1966.
22. Duncan, Otis Dudley, and Davis, Beverly. "An Alternative to Logical Correlation." American Sociological Review 18 (Dec. 1953): 665-666.
23. Duncan, Otis Dudley; Cuzzort, Ray; and Duncan, Beverly. Statistical Geography: Problems in Analyzing Areal Data. New York: The Free Press of Glencoe, 1961.
24. Edwards, J., and Orcutt, G. "Should Aggregation Prior to Estimation be the Rule?" Review of Economics and Statistics 51 (November 1969): 409-420.

25. Farkas, George. "Specification, Residuals and Contextual Effects." Sociological Methods and Research 2 (Feb. 1974): 333-363.
26. Feige, Edgar L., and Watts, Harold W. "An Investigation of the Consequences of Partical Aggregation of Micro-Economic Data." Econometrica 40 (Mar. 1972): 343-360.
27. Fennessy, James. "The General Linear Model: A New Perspective on Some Familiar Topics." American Journal of Sociology 74 (July 1968): 1-27.
28. Fleitas, Daniel. "Bandwagon and Underdog Effects in a Minimal Information Election." American Political Science Review 65 (June 1971): 434-438.
29. Goodman, Leo A. "Ecological Regressions and the Behavior of Individuals." American Sociological Review 18 (Dec. 1953): 663-664.
30. Goodman, Leo A. "Some Alternatives to Ecological Correlation." American Journal of Sociology 64 (May 1959): 610-625.
31. Green, H. A. John. Aggregation in Economic Analysis. Princeton: Princeton University Press, 1964.
32. Grunfeld, Yehuda, and Griliches, Zvi. "Is Aggregation Necessarily Bad?" Review of Economics and Statistics 42 (Feb. 1960): 1-13.
33. Gupta, Kanhaya Lai. Aggregation in Economics: A Theoretical and Empirical Study. Rotterdam: Rotterdam University Press, 1969.
34. Hammond, Jr., John L. "Two Sources of Error in Ecological Correlations." American Sociological Review 38 (Dec. 1973): 764-777.
35. Hannan, Michael T. Problems of Aggregation and Disaggregation in Sociological Research. Working papers in Methodology, No. 4, Chapel Hill: Institute for Research in Social Science, University of North Carolina, 1970.
36. Hannan, Michael, and Rurstein, Leigh. "Estimation from Grouped Observations." American Sociological Review 39 (June 1974): 374-392.

37. Hannan, Michael T.; Young, Alice A.; and Nielson, Francois. "Specification Bias Analysis of the Effects of Grouping of Observations in Multiple Regression Models." Paper presented at the Annual Meeting of the American Educational Research Association, April 1975, Washington, D.C.
38. Hanushek, Eric A.; Jackson, John E.; and Kain, John F. "Model Specification, Use of Aggregate Data, and the the Ecological Correlation Fallacy." Political Methodology (Winter 1974): 87-106.
39. Hanushek, Eric A., and Kain, John F. "On the Value of 'Equality of Educational Opportunity' As a Guide to Public Policy." Ch. 3 in Frederick Mosteller and Daniel P. Moynihan, eds., On Equality of Educational Opportunity. New York: Random House, 1972.
40. Harder, T., and Pappi, F. U. "Multiple-Level Regression Analysis of Survey and Ecological Data." Social Science Information 8 (Oct. 1969): 43-67.
41. Hauser, Robert M. "Context and Consequence: A Cautionary Tale." American Journal of Sociology 75 (Jan. 1970): 645-663.
42. Hauser, Robert M. "Contextual Analysis Revisited." Sociological Methods and Research 2 (Feb. 1974): 365-375.
43. Irwin, Laura. "Expenditure Effects of Federal Aid: Data Aggregation and the Risk of Uncertainty." Publius, Fall 1975.
44. Iversen, Gudman R. "Recovering Individual Data in the Presence of Group and Individual Effects." American Journal of Sociology 79 (1973): 420-434.
45. Jones, E. Terrence. "Ecological Inference and Electoral Analysis." The Journal of Interdisciplinary History 2 (1972): 249-269.
46. Kousser, J. Morgan. "Ecological Regression and the Analysis of Past Politics." Journal of Interdisciplinary History 4 (Autumn 1973): 237-262.
47. Lichtman, Alan J. "Correlation, Regression and the Ecological Fallacy: A Critique." The Journal of Interdisciplinary History 4 (Winter 1974): 417-433.

48. Juan J. Linz, "Ecological Analysis and Survey Research." In M. Dogan and S. Rokkan, eds., Quantitative Ecological Analysis in the Social Sciences, Cambridge: MIT Press, 1969, pp. 91-132.
49. McLaughlin, Milbrey Wallen. Evaluation and Reform: The Elementary and Secondary Education Act of 1965, Title I. Cambridge: Ballinger Publishing Co., 1975.
50. Meckstroth, Theodore W. "Some Problems in Cross-Level Inference." American Journal of Political Science 18 (Feb. 1974): 45-66.
51. Orcutt, Guy H. "Research Strategy on Modeling Economic Systems." In D. G. Watts, ed., The Future of Statistics. New York: Academic Press, 1968.
52. Orcutt, Guy H.; Watts, H.; and Edwards, Jr. "Data Aggregation and Information Loss." American Economic Review 58 (Sept. 1968): 773-787.
53. Peston, M. H. "A View of the Aggregation Problem." Review of Economic Studies 27 (1959/60): 58-64.
54. Porter, David O.; with Warner, D., and Porter, T. The Politics of Budgeting Federal Aid: Resource Mobilization by Local School Districts. Beverly Hills, California: Sage Publications, Inc., 1973.
55. Prais, S. J., and Aitchison, J. "The Grouping of Observations in Regression Analysis." Review of the International Statistical Institute 22 (1954): 1-22.
56. Przeworski, Adam. "Contextual Models of Political Behavior." Political Methodology 1 (Winter 1974): 27-60.
57. Robinson, William S. "Ecological Correlations and The Behavior of Individuals." American Sociological Review 15 (June 1950): 351-357.
58. Rusk, Jerrold G. "The Effect of the Australian Ballot Reform on Split Ticket Voting: 1876-1908." American Political Science Review 64 (Dec. 1970): 1220-1238.
59. Sacks, Seymour. City Schools/Suburban Schools: A History of Fiscal Conflict. Syracuse: Syracuse University Press, 1972.

60. Sacks, Seymour, and Harris, R. "The Determinants of State and Local Government Expenditures and Intergovernmental Flows of Funds." National Tax Journal 17 (March 1964): 75-85.
61. Salamon, Lester M., and Van Evera, Stephen. "Fear, Apathy, and Discrimination: A Test of Three Explanations of Political Participation." American Political Science Review 67 (Dec. 1973): 1288-1306.
62. Scheuch, Erwin K. "Cross-National Comparisons Using Aggregate Data: Some Substantive and Methodological Problems." In Comparing Nations, New Haven: Yale University Press, 1966, pp. 131-167.
63. Shively, W. Phillips. "Ecological Inference: The Use of Aggregate Data to Study Individuals." American Political Science Review 63 (Dec. 1969): 1183-1196.
64. Slatin, Gerald T. "Ecological Analysis of Delinquency: Aggregation Effects." American Sociological Review 34 (Dec. 1969): 894-907.
65. Soares, Glaucio, and Hamblin, Robert L. "Socioeconomic Variables and Voting for the Radical Left: Chile, 1952." American Political Science Review 61 (Dec. 1967): 1053-1065.
66. Stokes, Donald F. "Cross-Level Inference as a Game Against Nature." In J. L. Bernd, ed., Mathematical Applications in Political Science 4, Charlottesville: The University of Virginia Press, 1969, pp. 62-83.
67. Sullivan, John. "Multiple Indicators and Complex Causal Models." Ch. 18 in H. M. Blalock, ed., Causal Models in the Social Sciences, Chicago: Aldine-Atherton, 1971.
68. Tannenbaum, Arnold S., and Bachman, Jerald G. "Structural Versus Individual Effects." American Journal of Sociology 69 (1964): 585-595.
69. Theil, Henri. Linear Aggregation of Economic Relations. Amsterdam: North Holland Publishing Co., 1954.
70. Valkonen, Tapani. "Secondary Analysis of Survey Data with Ecological Variables." Social Science Information 8 (Dec. 1969): 33-36.

71. Walker, Jack. "The Diffusion of Innovations Among the American States." American Political Science Review 63 (Sept. 1969): 880-899.
72. Wilensky, Gail R. State Aid and Educational Opportunity. Beverly Hills, California: Sage Publications, 1970.